TRANSACTIONS

OF THE

AMERICAN INSTITUTE

OF

ELECTRICAL ENGINEERS

VOLUME 75

1956

PART II. APPLICATIONS AND INDUSTRY

PUBLISHED BY THE

AMERICAN INSTITUTE OF ELECTRICAL ENGINEERS

33 WEST 39TH STREET

NEW YORK 18, N. Y., U. S. A.

Copyright 1957 by the American Institute of Electrical Engineers

Printed in U. S. A. by The Mack Printing Company, Easton, Pa.

Preface

The AIEE *Transactions* for 1956 (volume 75) is published in three parts: Part I. Communication and Electronics; Part II. Applications and Industry; and Part III. Power Apparatus and Systems. The papers in each of the three parts are classified according to subject matter as follows:

Part I. Communication and Electronics	Part II. Applications and Industry	Part III. Power Apparatus and Systems
Communication Switching Systems	Air Transportation	Carrier Current
Communication Theory	Domestic and Commercial Applications	Insulated Conductors
Radio Communications Systems	Land Transportation	Power Generation
Special Communications Applications	Marine Transportation	Protective Devices
Telegraph Systems	Production and Application of Light	Relays
Television and Aural Broadcasting Systems	Chemical Industry	Rotating Machinery
Wire Communications Systems	Electric Heating	Substations
Basic Sciences	Electric Welding	Switchgear
Computing Devices	Feedback Control Systems	System Engineering
Dielectrics	General Industry Applications	Transformers
Electrical Techniques in Medicine and Biology	Industrial Control	Transmission and Distribution
Electronics	Industrial Power Rectifiers	
Instruments and Measurements	Industrial Power Systems	
Magnetic Amplifiers	Mining and Metal Industry	
Metallic Rectifiers	Petroleum Industry	
Nucleonics		Education
Solid State Devices		Safety
		Research

Each part has been indexed separately in the back of that particular part. The three parts are not cross-referenced; hence the user should determine first whether the subject matter of the paper desired is in the field of communication and electronics, applications and industry, or power apparatus and systems. Papers are listed in the subject index under several key words in the titles. The original numbers assigned to the papers are given in the author index. Volume 75 contains the technical papers and related discussions presented at these meetings:

1. Winter General Meeting, New York, N. Y., January 30–February 3, 1956.
2. South West District Meeting, Dallas, Tex., April 2–4, 1956.
3. Great Lakes District Meeting, Fort Wayne, Ind., April 16–18, 1956.
4. North Eastern District Meeting, Rochester, N. Y., May 2–4, 1956.
5. Summer and Pacific General Meeting, San Francisco, Calif., June 25–29, 1956.
6. Fall General Meeting, Chicago, Ill., October 1–5, 1956.

Statements and opinions given in the papers and discussions published in *Transactions* are the expressions of the contributors for which the American Institute of Electrical Engineers assumes no responsibility.

Applications and Industry Index for 1956

1. Technical Subject Index

A-C Electric Systems, Negative-Sequence Voltage Detection on Aircraft. Jessee, Boggess.....171–5
A-C Generators: Reactive Load Division and Selective Protection, Parallel Operation of Aircraft. Kahle.....175–80; disc. 180
A-C Mine Hoist Control System, A 2-Motor, Myles.....10–16
A-C Servo Systems, Envelope Transfer Function Analysis in. Panzer.....274–9
A-C System Problems, Analogue Computer Methods Applied to Steady State. Kahle, McConnell.279–82
Air-Break Contactor, Interruption Tests on High-Voltage. Lister.....56–9; disc. 60
Air-Cleaning Blower for Ventilating Traction Equipment, An. Gallagher.....159–61
Air-Frame Controller Designed by Phase-Space Analysis, A Study of a Predictor-Type. Hopkin, Iwama.....1–9
(Air Transportation) Analogue Computer Methods Applied to Steady-State A-C System Problems. Kahle, McConnell.....279–82
(Air Transportation) Analyzing Missile Electric System Reliability. Wilson.....206–12; disc. 213
(Air Transportation) Important Factors Influencing the Choice Between a Synchro and a Potentiometer as an Angular Position Pickoff. Dinter.....198–204; disc. 204
Aircraft A-C Electric Systems, Negative-Sequence Voltage Detection on. Jessee, Boggess.....171–5
Aircraft A-C Generators: Reactive Load Division and Selective Protection, Parallel Operation of. Kahle.....175–80; disc. 180
Aircraft Electric Systems, Development of Definitions for Voltage Modulation Characteristics in 380- to 420-Cycle-Per-Second. Markowitz.....193–7; disc. 197
Aircraft Electric Systems, Frequency Modulation and Load-Division Instability in 400-Cycle. Oman.....181–5; disc. 185
Aircraft Equipment Electric Insulation, High-Potential Testing of. Kilman, Dallas.....189–93; disc. 193
Aircraft Ground Power Electric Equipment, Recent Trends in. Sechrist, Rosenblatt, Bryant.....214–23; disc. 223
Analogue Computer Methods Applied to Steady State A-C System Problems. Kahle, McConnell.279–82
Analysis, A Study of a Predictor-Type Air-Frame Controller Designed by Phase-Space. Hopkin, Iwama.....1–9
Analysis and Performance of a Valve-Controlled Hydraulic Servomechanism. Stallard.....75–83; disc. 83
Analysis by Direct Area Measurements, Harmonic. Dortort.....16–19
Analysis in A-C Servo Systems, Envelope Transfer Function. Panzer.....274–9
Analysis of Feedback Systems with One Nonlinear Element, A Step-by-Step Method for Transient. Stout.....378–89; disc. 389
Analysis of Servomechanisms with Nonlinear Feedback Control. Ku.....402–05; disc. 405
Analysis of Tandem Cold Reduction Mill with Automatic Gauge Control. Phillips.....355–63; disc. 370
Analysis of Time-Varying Sampled-Data Systems, A Technique for the. Friedland.....407–11; disc. 411
Analysing Missile Electric System Reliability. Wilson.....206–12; disc. 213
Angular Position Pickoff, Important Factors Influencing the Choice Between a Synchro and a Potentiometer as an. Dinter.....198–204; disc. 204
Applying Industrial Controllers for Safe Operation. Heumann.....291–5
Arc Behavior and Metal Transfer Characteristics, Inert Gas-Shielded Welding. Skinner, Yenni.....28–32
Arc-Drop Voltage on Mercury-Arc Rectifiers, Methods of Measuring. Winograd, Lawton.....43–7; disc. 47
Arcs, Influence of Atmospheric Water Vapor on High-Current D-C. Benner, Jones.....162–6
Argon Atmospheres, Cathode Instability in. Ludwig.....32–5
Asymptotes, A General Technique for Approximating Transient Response from Frequency Response. Biernson.....253–73
Atmospheres, Cathode Instability in Argon. Ludwig.....32–5
Atmospheric Water Vapor on High Current D-C Arcs, Influence of. Benner, Jones.....162–6
Automatic Control System, Describing Functions Representing the Effect of Inertia, Backlash, and Coulomb Friction on the Stability of an. Satyendra. Part I.....243–8; disc. 248
Automatic Output Regulator for Power Oscillators, An. Hopkins, Fraser.....36–8

B

Backlash, and Coulomb Friction on the Stability of an Automatic Control System, Describing Functions Representing the Effect of Inertia. Satyendra. Part I.....243–8; disc. 248
Backlash and Hysteresis Effects in Second-Order Feedback Systems, Comparison of. Vallese.....240–3
Block Diagram Transformations for Systems with One Nonlinear Element. Stout.....130–9; disc. 139
Blower for Ventilating Traction Equipment, An Air-Cleaning. Gallagher.....159–61

C

Cathode Circuit Breaker, A New High-Power. McConnell, Findley.....39–42; disc. 42
Cathode Instability in Argon Atmospheres. Ludwig.....32–5
Centralized Control of Power Systems in the Steel Industry. Derr, Metz.....332–7; disc. 337
Characteristics of D-C Metal-Rolling Motors and Generators, Transient. Smith.....326–31; disc. 331
(Chemical Industry) Electric System for a Nuclear Reactor. Peabody, Brereton.....296–306
Circuit Breaker, A New High-Power Cathode. McConnell, Findley.....39–42; disc. 42
Circuits for Industrial Control Systems, Magnetic Logic. Evans, Hall, Van Nice.....166–71
Cold Reduction Mill with Automatic Gauge Control, Analysis of Tandem. Phillips.....355–63; disc. 370
Commercial-Frequency Railroad Electrification, Power Supply for. Price.....441
Comparison of Backlash and Hysteresis Effects in Second-Order Feedback Systems. Vallese.....240–3
Compensating Network, A Time-Dependent Nonlinear. Clegg.....306–08; disc. 308
Compensation of Continuous-Data Feedback Control Systems, Digital. Maitra, Sarachik.....107–14; disc. 115
Compensation of Sampled-Data and Continuous Control Systems, Discrete. Jury, Schroeder.....317–25
Computer Methods Applied to Steady State A-C System Problems, Analogue. Kahle, McConnell.....279–82
Contactor, Interruption Tests on High-Voltage Air-Break. Lister.....56–9; disc. 60
Contactor Servo Using Describing Function Theory, Design of a. Chao.....223–31; disc. 231
Contour and Root-Contour Methods, Synthesis of Feedback Control Systems by Gain-. Yeh.....85–96; disc. 96
Control of Power Systems in the Steel Industry, Centralized. Derr, Metz.....332–7; disc. 337
Control Switch with a Pulsation Filter, A Pressure. Schaefer.....426–32
Control System to Random Signals, Response of a Particular Nonlinear. Nikiforuk.....419–22
Control Systems, Discrete Compensation of Sampled-Data and Continuous. Jury, Schroeder.....317–25
Control Systems, Hidden Oscillations in Sampled-Data. Jury.....391–5
Control Systems, Synthesis and Critical Study of Sampled-Data. Jury.....141–9; disc. 149
Controller Designed by Phase-Space Analysis, A Study of a Predictor-Type Air-Frame. Hopkin, Iwama.....1–9
Controllers for Safe Operation, Applying Industrial. Heumann.....291–5
Controllers for Sampled-Data Feedback Systems, Factors in the Design of Digital. Bertram.....151–8; disc. 158
Coulomb Friction on the Stability of an Automatic Control System, Describing Functions Representing the Effect of Inertia, Backlash, and. Satyendra. Part I.....243–8; disc. 248
Coulomb, Static, and Viscous Friction, A Note on the Describing Function of an Element with. Silberberg.....423–5
Currents and Voltages, Let-Go. Dalziel, Massoglia.....49–55; disc. 55

D

D-C Arcs, Influence of Atmospheric Water Vapor on High-Current. Benner, Jones.....162–6
D-C Metal-Rolling Motors and Generators, Transient Characteristics of. Smith.....326–31; disc. 331
Damped Servomechanisms, Variably. Schieber.....414–48; disc. 419
Data Feedback Control Systems, Digital Compensation of Continuous- Maitra, Sasachik.....107–14; disc. 115
Definitions for Voltage Modulation Characteristics in 380- to 420-Cycle-Per-Second Aircraft Electric Systems, Development of. Markowitz.....193–7; disc. 197
Describing Function of an Element with Coulomb, Static, and Viscous Friction, A Note on the. Silberberg.....423–5
Describing Function Theory, Design of a Contactor Servo Using. Chao.....223–31; disc. 231
Describing Functions Representing the Effect of Inertia, Backlash, and Coulomb Friction on the Stability of an Automatic Control System. Satyendra. Part I.....243–8; disc. 248
Design of a Contactor Servo Using Describing Function Theory. Chao.....223–31; disc. 231
Detection on Aircraft A-C Electric Systems, Negative-Sequence Voltage. Jessee, Boggess.....171–5
Development of Definitions for Voltage Modulation Characteristics in 380- to 420-Cycle-Per-Second Aircraft Electric Systems. Markowitz.....193–7; disc. 197
Diagram Transformations for Systems with One Nonlinear Element, Block. Stout.....130–9; disc. 139
Digital Compensation of Continuous-Data Feedback Control Systems. Maitra, Sarachik.....107–14; disc. 115
Digital Controllers for Sampled-Data Feedback Systems, Factors in the Design of. Bertram.....151–8; disc. 158
Dips and Flicker, Voltage. Kronenberg.....349–51; disc. 351
Direct Area Measurements, Harmonic Analysis by. Dortort.....16–19
Discrete Compensation of Sampled-Data and Continuous Control Systems. Jury, Schroeder.....317–25
Distributing Large Single-Phase Loads in Limited Capacity 3-Phase Power Systems. Aroyan, Lakey.....249–52

E

Economic Studies Applied to Industrial Power Systems Backenstoss.....61–7; disc. 67
Electric Equipment, Recent Trends in Aircraft Ground Power. Sechrist, Rosenblatt, Bryant.....214–23; disc. 223
(Electric Heating) An Automatic Output Regulator for Power Oscillators. Hopkins, Fraser.....36–8
Electric Insulation, High-Potential Testing of Aircraft Equipment. Kilman, Dallas.....189–93; disc. 193
Electric Power, Shipboard Use of 400-Cycle. Apple, Lusby.....116–21; disc. 121
Electric Power Systems for Steel Plants. Cox, Levoy.....339–48; disc. 348
Electric System for a Nuclear Reactor. Peabody, Brereton.....296–306
Electric System Reliability, Analyzing Missile. Wilson.....206–12; disc. 213
Electric Systems, Negative-Sequence Voltage Detection on Aircraft A-C. Jessee, Boggess.....171–5
(Electric Welding) Cathode Instability in Argon Atmospheres. Ludwig.....32–5
(Electric Welding) Influence of Atmospheric Water Vapor on High Current D-C Arcs. Benner, Jones.....162–6
(Electric Welding) Let-Go Currents and Voltages. Dalziel, Massoglia.....49–55; disc. 55
Element, A Step-by-Step Method for Transient Analysis of Feedback Systems with One Nonlinear. Stout.....378–89; disc. 389
Elements with Coulomb, Static, and Viscous Friction, A Note on the Describing Function of an. Silberberg.....423–5
Envelope Transfer Function Analysis in A-C Servo Systems. Panzer.....274–9
Equations for Tandem Rolling Mills, Incremental Control. Courcoulas, Ham.....363–70; disc. 371
Equipment Electric Insulation, High-Potential Testing of Aircraft. Kilman, Dallas.....189–93; disc. 193

Equipment, Recent Trends in Aircraft Ground Power Electric. Sechrist, Rosenblatt, Bryant. . . . . 214–23; disc. 223

Evaluation of Radio Influence Voltages in Fluorescent Lighting Installations. Wright, Zimmerman. . . . . 96–9; disc. 99

Experimental Treatment of Nonlinear Servomechanisms, An. Neiswander. 308–16

F

Factors in the Design of Digital Controllers for Sampled-Data Feedback Systems. Bertram. . . . . 151–8; disc. 158

Factors Influencing the Choice Between a Synchro and a Potentiometer as an Angular Position Pickoff, Important. Dinter. . . . . 198–204; disc. 204

(Feedback Control) A General Technique for Approximating Transient Response from Frequency Response Asymptotes. Biernson. 253–73

(Feedback Control) A Note on the Describing Function of an Element with Coulomb, Static, and Viscous Friction. Silberberg. 423–5

(Feedback Control) A Study of a Predictor-Type Air-Frame Controller Designed by Phase-Space Analysis. Hopkin, Iwama. 1–9

(Feedback Control) A Technique for the Analysis of Time-Varying Sampled-Data Systems. Friedland. . . . . 407–11; disc. 411

(Feedback Control) Analysis and Performance of a Valve-Controlled Hydraulic Servomechanism. Stallard. . . . . 75–83; disc. 83

Feedback Control, Analysis of Servomechanisms with Nonlinear. Ku. . . . . 402–05; disc. 405

(Feedback Control) Block Diagram Transformations for Systems with One Nonlinear Element. Stout. . . . . 130–9; disc. 139

(Feedback Control) Describing Functions Representing the Effect of Inertia, Backlash, and Coulomb Friction on the Stability of an Automatic Control System. Satyendra. Part I. . . . . 243–8; disc. 248

(Feedback Control) Design of a Contactor Servo Using Describing Function Theory. Chao. . . . . 223–31; disc. 231

(Feedback Control) Discrete Compensation of Sampled-Data and Continuous Control Systems. Jury, Schroeder. 317–25

(Feedback Control) Envelope Transfer Function Analysis in A-C Servo Systems. Panser. 274–9

(Feedback Control) Hidden Oscillations in Sampled-Data Control Systems. Jury. 391–5

(Feedback Control) Incremental Control Equations for Tandem Rolling Mills. Courcoulas, Ham. . . . . 363–70; disc. 371

(Feedback Control) Response of a Particular Nonlinear Control System to Random Signals. Nikiforuk. 419–22

(Feedback Control) Synthesis and Critical Study of Sampled-Data Control Systems. Jury. . . . . 141–9, disc. 149

Feedback Control Systems by Gain-Contour and Root-Contour Methods, Synthesis of. Yeh. . . . . 85–96; disc. 96

Feedback Control Systems, Digital Compensation of Continuous-Data. Maitra, Sarachik. . . . . 107–14; disc. 115

(Feedback Control) The Frequency Response of a Servomechanism Designed for Optimum Transient Response. West, Nikiforuk. 234–9

(Feedback Control) Variably Damped Servomechanisms. Schieber. . . . . 414–18; disc. 419

Feedback Systems, Comparison of Backlash and Hysteresis Effects in Second-Order. Vallese. 240–3

Feedback Systems, Factors in Design of Digital Controllers for Sampled-Data. Bertram. 151–8; disc. 158

Feedback Systems with One Nonlinear Element, A Step-by-Step Method for Transient Analysis of. Stout. . . . . 378–89; disc. 389

Feedback Techniques Applied to the Single-Phase Transformer. Mishkin. . . . . 374–6; disc. 376

Filter, A Pressure Control Switch with a Pulsation. Schaefer. 426–32

Flicker, Voltage Dips and. Kroneberg. 349–51; disc. 351

Fluorescent Lighting Installations, Evaluation of Radio Influence Voltages in. Wright, Zimmermann. . . . . 96–9; disc. 99

400-Cycle Aircraft Electric Systems, Frequency Modulation and Load-Division Instability in. Oman. . . . . 181–5; disc. 185

400-Cycle Electric Power, Shipboard Use of. Apple, Lusby. . . . . 116–21; disc. 121

Frequency Modulation and Load-Division Instability in 400-Cycle Aircraft Electric Systems. Oman. . . . . 181–5; disc. 185

Frequency Railroad Electrification, Power Supply for Commercial-. Price. . . . . 432–41; disc. 441

Frequency Response Asymptotes, A General Technique for Approximating Transient Response from. Biernson. 253–73

Frequency Response of a Servomechanism Designed for Optimum Transient Response, The. West, Nikiforuk. 234–9

Friction, A Note on the Describing Function of an Element with Coulomb, Static, and Viscous. Silberberg. 423–5

Friction on the Stability of an Automatic Control System, Describing Functions Representing the Effect of Inertia, Backlash, and Coulomb. Satyendra. Part I. . . . . 243–8; disc. 248

Function Analysis in A-C Servo Systems, Envelope Transfer. Panser. 274–9

Function of an Element with Coulomb, Static, and Viscous Friction, A Note on the Describing. Silberberg. 423–5

Function Theory, Design of a Contactor Servo Using Describing. Chao. . . . . 223–31; disc. 231

Functions Representing the Effect of Inertia, Backlash, and Coulomb Friction on the Stability of an Automatic Control System, Describing. Satyendra. Part I. . . . . 243–8; disc. 248

G

Gain-Contour and Root-Contour Methods, Synthesis of Feedback Control Systems by. Yeh. 85–96; disc. 96

Gas-Shielded Welding Arc Behavior and Metal Transfer Characteristics, Inert. Skinner, Yenni. 28–32

Gauge Control, Analysis of Tandem Cold Reduction Mill with Automatic. Phillips. . . . . 355–63; disc. 370

Generators. Reactive Load Division and Selective Protection, Parallel Operation of Aircraft A-C. Kahle. . . . . 175–80; disc. 180

Generators, Transient Characteristics of D-C Metal-Rolling Motors and. Smith. . . . . 326–31; disc. 331

Ground Power Electric Equipment, Recent Trends in Aircraft. Sechrist, Rosenblatt, Bryant. . . . . 214–23; disc. 223

H

Harmonic Analysis by Direct Area Measurements. Dortort. 16–19

Hidden Oscillations in Sampled-Data Control Systems. Jury. 391–5

High-Potential Testing of Aircraft Equipment Electric Insulation. Kilman, Dallas. . . . . 189–93; disc. 193

High-Voltage Air-Break Contactor, Interruption Tests on. Lister. . . . . 56–9; disc. 60

Hoist Control System, A 2-Motor A-C Mine. Myles. 10–16

Hydraulic Servomechanism, Analysis and Performance of a Valve-Controlled. Stallard. . . . . 75–83; disc. 83

Hysteresis Effects in Second-Order Feedback Systems, Comparison of Backlash and. Vallese. 240–3

I

Ignitron Principles Extended to Larger Tubes, Sealed. Boyer, Colaiaco. . . . . 125–8; disc. 128

Important Factors Influencing the Choice Between a Synchro and a Potentiometer as an Angular Position Pickoff. Dinter. . . . . 198–204; disc. 204

Incremental Control Equations for Tandem Rolling Mills. Courcoulas, Ham. . . . . 363–70; disc. 371

(Industrial Control) A Pressure Control Switch with a Pulsation Filter. Schaefer. 426–32

(Industrial Control) A 2-Motor A-C Mine Hoist Control System. Myles. 10–16

(Industrial Control) Interruption Tests on High-Voltage Air-Break Contactor. Lister. . . . . 56–9; disc. 60

Industrial Control Systems, Magnetic Logic Circuits for. Evans, Hall, Van Nice. 166–71

Industrial Controllers for Safe Operation, Applying. Heumann. 291–5

(Industrial Power Rectifiers) A New High-Power Cathode Circuit Breaker. McConnell, Findley. . . . . 39–42; disc. 42

(Industrial Power Rectifiers) Harmonic Analysis by Direct Area Measurements. Dortort. 16–19

(Industrial Power Rectifiers) Sealed Ignitron Principles Extended to Larger Tubes. Boyer, Colaiaco. . . . . 125–8; disc. 128

(Industrial Power Rectifiers) Test and Operating Experience with the Triple-Diametric Rectifier. Wechter, Hague, Marcum. . . . . 100–04; disc. 104

Industrial Power Systems, Economic Studies Applied to. Backenstoss. . . . . 61–7; disc. 67

(Industrial Power Systems) Voltage Dips and Flicker. Kroneberg. . . . . 349–51; disc. 351

Industry, Centralized Control of Power Systems in the Steel. Derr, Metz. . . . . 332–7; disc. 337

Inert Gas-Shielded Welding Arc Behavior and Metal Transfer Characteristics. Skinner, Yenni. . 28–32

Inertia, Backlash, and Coulomb Friction on the Stability of an Automatic Control System, Describing Functions Representing the Effect of. Satyendra. Part I. . . . . 243–8; disc. 248

Influence of Atmospheric Water Vapor on High-Current D-C Arcs. Benner, Jones. 162–6

Instability in 400-Cycle Aircraft Electric Systems, Frequency Modulation and Load-Division. Oman. . . . . 181–5; disc. 185

Insulation, High-Potential Testing of Aircraft Equipment Electric. Kilman, Dallas. 189–93; disc. 193

Interruption Tests on High-Voltage Air-Break Contactor. Lister. . . . . 56–9, disc. 60

L

(Land Transportation) An Air-Cleaning Blower for Ventilating Traction Equipment. Gallagher. 159–61

(Land Transportation) Modernization of the Long Island Rail Road Passenger-Car Fleet. Hatch. . . 395–401

(Land Transportation) Power Supply for Commercial-Frequency Railroad Electrification. Price. . . . . 432–41; disc. 441

(Land Transportation) Rapid Transit Progress in Toronto. Inglis. 352–4

Let-Go Currents and Voltages. Dalziel, Massoglia. . . . . 49–55; disc. 55

Lighting Installations, Evaluation of Radio Influence Voltages in Fluorescent. Wright, Zimmermann. . . . . 96–9; disc. 99

Limited-Capacity 3-Phase Power Systems, Distributing Large Single-Phase Loads in. Aroyan, Lakey. 249–52

Load-Division Instability in 400-Cycle Aircraft Electric Systems, Frequency Modulation and. Oman. . . . . 181–5; disc. 185

Loads in Limited-Capacity 3-Phase Power Systems, Distributing Large Single-Phase. Aroyan, Lakey. 294–52

Logic Circuits for Industrial Control Systems, Magnetic. Evans, Hall, Van Nice. 166–71

Long Island Railroad Passenger-Car Fleet, Modernization of the. Hatch. 395–401

M

Magnetic Logic Circuits for Industrial Control Systems. Evans, Hall, Van Nice. 166–71

(Marine Transportation) Shipboard Use of 400-Cycle Electric Power. Apple, Lusby. . . 116–21; disc. 121

Measurements, Harmonic Analysis by Direct Area. Dortort. 16–19

Measuring Arc-Drop Voltage on Mercury-Arc Rectifiers, Methods of. Winograd, Lawton. . . 43–7; disc. 47

Mercury-Arc Rectifiers, Methods of Measuring Arc-Drop Voltage on. Winograd, Lawton. . . . . 43–7; disc. 47

Metal-Rolling Motors and Generators, Transient Characteristics of D-C. Smith. . . . . 326–31; disc. 331

Metal Transfer Characteristics, Inert Gas-Shielded Welding Arc Behavior and. Skinner, Yenni. 28–32

Method for the Synthesis of Nonlinear Servomechanisms, A Phase-Space. Kuba, Kazda. . . . . 282–9; disc. 289

Method for Transient Analysis of Feedback Systems with One Nonlinear Element, A Step-by-Step. Stout. . . . . 378–89; disc. 389

Methods of Measuring Arc-Drop Voltage on Mercury-Arc Rectifiers. Winograd, Lawton. . . 43–7; disc. 47

Mill, Rectifier Power Supply for a Modern Rod. Buchanan, Hodgson. . . . . 70–4; disc. 74

Mill with Automatic Gauge Control, Analysis of Tandem Cold Reduction. Phillips. . . . . 355–63; disc. 370

Mills, Incremental Control Equations for Tandem Rolling. Courcoulas, Ham. . . . . 363–70; disc. 371

Mine Hoist Control System, A 2-Motor A-C. Myles. 10–16

(Mining and Metal Industry) Analysis of Tandem Cold Reduction Mill with Automatic Gauge Control. Phillips. . . . . 355–63; disc. 370

(Mining and Metal Industry) Centralized Control of Power Systems in the Steel Industry. Derr, Metz. . . . . 332–7; disc. 337

(Mining and Metal Industry) Electric Power Systems for Steel Plants. Cox, Levoy. . . . . 339–48, disc. 348

Missile Electric System Reliability, Analyzing. Wilson. . . . . 206–12; disc. 213

Modernization of the Long Island Rail Road Passenger Car Fleet. Hatch. 395–401

Motor A-C Mine Hoist Control System, A 2-Two-. Myles. 10–16

Motors and Generators, Transient Characteristics of D-C Metal-Rolling. Smith. . . . . 326–31; disc. 331

N

Negative-Sequence Voltage Detection on Aircraft A-C Electric Systems. Jessee, Borgess. 171–5

Network, A Time-Dependent Nonlinear Compensating. Clegg. . . . . 306–08; disc. 308

New High-Power Cathode Circuit Breaker, A. McConnell, Findley. . . . . 39–42; disc. 42

Nonlinear Compensating Network, A Time-Dependent. Clegg. . . . . 306–08; disc. 308

Nonlinear Control System to Random Signals, Response of a Particular. Nikiforuk. 419–22

Nonlinear Element, A Step-by-Step Method for Transient Analysis of Feedback Systems with One. Stout. . . . . 378–89; disc. 389

Nonlinear Element, Block Diagram Transformations for Systems with One. Stout....130–9; disc. 139
Nonlinear Feedback Control, Analysis of Servomechanisms with. Ku....402–05; disc. 405
Nonlinear Servomechanisms, A Phase-Space Method for the Synthesis of. Kuba, Kazda....282–9; disc. 289
Nonlinear Servomechanisms, An Experimental Treatment of. Neiwander....308–16
Note on the Describing Function of an Element with Coulomb, Static, and Viscous Friction, A. Silberberg....423–5
Nuclear Reactor, Electric System for a. Peabody, Brereton....296–306

O

Operating Experience with the Triple-Diametric Rectifier, Test and. Wachter, Hague, Marcum....100–04; disc. 104
Operation of Aircraft A-C Generators: Reactive Load Division and Selective Protection, Parallel. Kahle....175–80; disc. 180
Optimum Transient Response, The Frequency Response of a Servomechanism Designed for. West, Nikiforuk....234–9
Oscillations in Sampled-Data Control Systems, Hidden. Jury....391–5
Oscillators, An Automatic Output Regulator for Power. Hopkins, Fraser....36–8
Output Regulator for Power Oscillators, An Automatic. Hopkins, Fraser....36–8

P

Parallel Operation of Aircraft A-C Generators: Reactive Load Division and Selective Protection. Kahle....175–80; disc. 180
Passenger-Car Fleet, Modernization of the Long Island Rail Road. Hatch....395–401
Phase-Space Analysis, A Study of a Predictor-Type Air-Frame Controller Designed by. Hopkin, Iwama....1–9
Phase-Space Method for the Synthesis of Nonlinear Servomechanisms, A. Kuba, Kazda....282–9; disc. 289
Pickoff, Important Factors Influencing the Choice Between a Synchro and a Potentiometer as an Angular Position. Dinter....198–204; disc. 204
Potentiometer as an Angular Position Pickoff, Important Factors Influencing the Choice Between a Synchro and a. Dinter....198–204; disc. 204
Power Electric Equipment, Recent Trends in Aircraft Ground. Sechrist, Rosenblatt, Bryant....214–23; disc. 223
Power Oscillators, An Automatic Output Regulator for. Hopkins, Fraser....36–8
Power Supply for a Modern Rod Mill, Rectifier. Buchanan, Hodgson....70–4; disc. 74
Power Supply for Commercial-Frequecy Railroad Electrification. Price....432–41; disc. 441
Power Systems, Distributing Single-Phase Loads in Limited-Capacity 3-Phase. Aroyan, Lakey....294–52
Power Systems for Steel Plants, Electric. Cox, Levoy....339–48; disc. 348
Power Systems in the Steel Industry, Centralized Control of. Derr, Metz....332–7; disc. 337
Predictor-Type Air-Frame Controller Designed by Phase-Space Analysis, A Study of a. Hopkin, Iwama....1–9
Pressure Control Switch with a Pulsation Filter, A. Schaefer....426–32
Principles Extended to Larger Tubes, Sealed Ignitron. Boyer, Colalaco....125–8; disc. 128
Pulsation Filter, A Pressure Control Switch with a. Schaefer....426–32

R

Radio Influence Voltages in Fluorescent Lighting Installations, Evaluation of. Wright, Zimmermann....96–9; disc. 99
Railroad Electrification, Power Supply for Commercial-Frequency. Price....432–41; disc. 441
Random Signals, Response of a Particular Nonlinear Control System to. Nikiforuk....419–22
Rapid Transit Progress in Toronto. Inglis....352–4
Reactor, Electric System for a Nuclear. Peabody, Brereton....296–306
Recent Trends in Aircraft Ground Power Electric Equipment. Sechrist, Rosenblatt, Bryant....214–23; disc. 223
Rectifier Power Supply for a Modern Rod Mill. Buchanan, Hodgson....70–4; disc. 74
Rectifier, Test and Operating Experience with the Triple-Diametric. Wachter, Hague, Marcum....100–04; disc. 104
Rectifiers, Methods of Measuring Arc-Drop Voltage on Mercury-Arc. Winograd, Lawton....43–7; disc. 47
Regulator for Power Oscillators, An Automatic Output. Hopkins, Fraser....36–8
Reliability, Analyzing Missile Electric System. Wilson....206–12; disc. 213
Response from Frequency Response Asymptotes, A General Technique for Approximating Transient. Bierson....253–73
Response of a Particular Nonlinear Control System to Random Signals. Nikiforuk....419–22
Response of a Servomechanism Designed for Optimum Transient Response, The Frequency. West, Nikiforuk....234–9
Rod Mill, Rectifier Power Supply for a Modern. Buchanan, Hodgson....70–4; disc. 74
Rolling Mills, Incremental Control Equations for Tandem. Courcoulas, Ham....363–70; disc. 371
Root-Contour Methods, Synthesis of Feedback Control Systems by Gain-Contour and. Yeh..85–96; disc. 96

S

Safe Operation, Applying Industrial Controllers for. Heumann....291–5
Sampled-Data and Continuous Control Systems, Discrete Compensation of. Jury, Schroeder....317–25
Sampled-Data Control Systems, Hidden Oscillations in. Jury....391–5
Sampled-Data Control Systems, Synthesis and Critical Study of. Jury....141–9; disc. 149
Sampled-Data Feedback Systems, Factors in the Design of Digital Controllers for. Bertram....151–8; disc. 158
Sampled-Data Systems, A Technique for the Analysis of Time-Varying. Friedland....407–11; disc. 411
Sealed Ignitron Principles Extended to Larger Tubes. Boyer, Colalaco....125–8; disc. 128
Second-Order Feedback Systems, Comparison of Backlash and Hysteresis Effects in. Vallese....240–3
Servo Systems, Envelope Transfer Function Analysis in A-C. Panzer....274–9
Servo Using Describing Function Theory, Design of a Contactor. Chao....223–31; disc. 231
Servomechanism, Analysis and Performance of a Valve-Controlled Hydraulic. Stallard....75–83; disc. 83
Servomechanism Designed for Optimum Transient Response, The Frequency Response of a. West, Nikiforuk....234–9
Servomechanism, A Phase-Space Method for the Synthesis of Nonlinear. Kuba, Kazda. 282–9; disc. 289
Servomechanisms, An Experimental Treatment of Nonlinear. Neiwander....308–16
Servomechanisms, Two Types of Zero-Velocity-Error Levinson....19–27
Servomechanisms, Variably Damped. Schiebel....414–18; disc. 419
Servomechanisms with Nonlinear Feedback Control, Analysis of. Ku....402–05; disc. 405,
Shipboard Use of 400-Cycle Electric Power. Apple, Lusby....116–21; disc. 121
Signals, Response of a Particular Nonlinear Control System to Random. Nikiforuk....419–22
Single-Phase Loads in Limited Capacity 3-Phase Power Systems, Distributing Large. Aroyan, Lakey....249–52
Single-Phase Transformer, Feedback Techniques Applied to the. Minkin....374–6; disc. 376
Static, and Viscous Friction, A Note on the Describing Function of an Element with Coulomb. Silberberg....423–5
Steady State A-C System Problems, Analogue Computer Methods Applied to. Kahle, McConnell....279–82
Steel Industry, Centralized Control of Power Systems in the. Derr, Metz....332–7; disc. 337
Steel Plants, Electric Power Systems for. Cox, Levoy....339–48; disc. 348
Step-by-Step Method for Transient Analysis of Feedback Systems with One Nonlinear Element, A. Stout....378–89; disc. 389
Studies Applied to Industrial Power Systems, Economic. Backenstos....61–7; disc. 67
Study of a Predictor-Type Air-Frame Controller Designed by Phase-Space Analysis, A. Hopkin, Iwama....1–9
Study of Sampled-Data Control Systems, Synthesis and Critical. Jury....141–9; disc. 149
(Subway) Rapid Transit Progress in Toronto. Inglis....352–4
Switch with a Pulsation Filter, A Pressure Control. Schaefer....426–32
Synchro and a Potentiometer as an Angular Position Pickoff, Important Factors Influencing the Choice Between a. Dinter....198–204; disc. 204
Synthesis and Critical Study of Sampled-Data Control Systems. Jury....141–9; disc. 149
Synthesis of Feedback Control Systems by Gain-Contour and Root-Contour Methods. Yeh..85–96; disc. 96
Synthesis of Nonlinear Servomechanisms, A Phase-Space Method for the. Kuba, Kazda....282–9; disc. 289
System Problems, Analogue Computer Methods Applied to Steady-State A-C. Kahle, McConnell...279–82
System Reliability, Analyzing Missile Electric. Wilson....206–12; disc. 213
Systems, A Technique for the Analysis of Time-Varying Sampled-Data. Friedland....407–11; disc. 411
Systems, Hidden Oscillations in Sampled-Data Control. Jury....391–5
Systems, Magnetic Logic Circuits for Industrial Control Evans, Hall, Van Nice....166–71.

T

Tandem Cold Reduction Mill with Automatic Gauge Control, Analysis of. Phillips....355–63; disc. 370
Tandem Rolling Mills, Incremental Control Equations for. Courcoulas, Ham....363–70; disc. 371
Technique for Approximating Transient Response from Frequency Response Asymptotes, A General. Bierson....253–73
Technique for the Analysis of Time-Varying Sampled-Data Systems, A. Friedland....407–11; disc. 411
Test and Operating Experience with the Triple-Diametric Rectifier. Wachter, Hague, Marcum....100–04; disc. 104
Testing of Aircraft Equipment Electric Insulation, High-Potential. Kilman, Dallas....189–93; disc. 193
Tests on High-Voltage Air-Break Contactor, Interruption. Lister....56–9; disc. 60
Theory, Design of a Contactor Servo Using Describing Function. Chao....223–31; disc. 231
380- to 420-Cycle-Per-Second Aircraft Electric Systems, Development of Definitions for Voltage Modulation Characteristics in. Markowitz....193–7; disc. 197
Three-Phase Power Systems, Distributing Large Single-Phase Loads in Limited Capacity. Aroyan, Lakey....249–52
Time-Dependent Nonlinear Compensating Network, A. Clegg....306–08; disc. 308
Time-Varying Sampled-Data Systems, A Technique for the Analysis of. Friedland....407–11; disc. 411
Toronto, Rapid Transit Progress in. Inglis....352–4
Traction Equipment, An Air-Cleaning Blower for Ventilating. Gallagher....159–61
Transfer Characteristics, Inert Gas-Shielded Welding Arc Behavior and Metal. Skinner, Yenni....28–32
Transfer Function Analysis in A-C Servo Systems, Envelope. Panzer....274–9
Transformations for Systems with One Nonlinear Element, Block Diagram. Stout....130–9; disc. 139
Transformer, Feedback Techniques Applied to the Single-Phase. Minkin....374–6; disc. 376
Transient Analysis of Feedback Systems with One Nonlinear Element, A Step-by-Step Method for. Stout....378–89; disc. 389
Transient Characteristics of D-C Metal-Rolling Motors and Generators. Smith....326–31; disc. 331
Transient Response from Frequency Response Asymptotes, A General Technique for Approximating. Bierson....253–73
Transient Response, The Frequency Response of a Servomechanism Designed for Optimum. West, Nikiforuk....234–9
(Transportation, Land) An Air-Cleaning Blower for Ventilating Traction Equipment. Gallagher....159–61
(Transportation, Land) Modernization of the Long Island Rail Road Passenger-Car Fleet. Hatch....395–401
(Transportation, Land) Power Supply for Commercial-Frequency Railroad Electrification. Price....432–41; disc. 441
(Transportation, Land) Rapid Transit Progress in Toronto. Inglis....352–4
Treatment of Nonlinear Servomechanisms, An Experimental. Neiwander....308–16
Trends in Aircraft Ground Power Electric Equipment, Recent. Sechrist, Rosenblatt, Bryant....214–23; disc. 223
Triple-Diametric Rectifier, Test and Operating Experience with the. Wachter, Hague, Marcum....100–04; disc. 104
Tubes, Sealed Ignitron Principles Extended to Larger. Boyer, Colalaco....125–8; disc. 128
Two-Motor A-C Mine Hoist Control System, A. Hylne....10–16
Two Types of Zero-Velocity-Error Servomechanisms, Levinson....19–27.

V

Valve-Controlled Hydraulic Servomechanism, Analysis and Performance of a. Stallard....75–83; disc. 83
Vapor on High-Current D-C Arcs, Influence of Atmospheric Water. Benner, Jones....162–6
Variably Damped Servomechanisms. Schiebel....414–18; disc. 419
Velocity-Error Servomechanisms, Two Types of Zero-. Levinson....19–27
Ventilating Traction Equipment, An Air-Cleaning Blower for. Gallagher....159–61
Voltage Detection on Aircraft A-C Electric Systems Negative-Sequence. Jones, Boggess....171–5
Voltage Dips and Flicker. Kroneberg....349–51; disc. 351

Voltage Modulation Characteristics in 380- to 420-Cycle-Per-Second Aircraft Electric Systems, Development of Definitions for. Markowitz....193–7; disc. 197
Voltage on Mercury-Arc Rectifiers, Methods of Measuring Arc-Drop. Winograd, Lawton...43–7; disc. 47
Voltages in Fluorescent Lighting Installations, Evaluation of Radio Influence. Wright, Zimmermann...96–9; disc. 99
Voltages, Let-Go Currents and. Dalziel, Massoglia...49–55; disc. 55

W

Water Vapor on High-Current D-C Arcs, Influence of Atmospheric. Benner, Jones....162–6
Welding Arc Behavior and Metal Transfer Characteristics, Inert Gas-Shielded. Skinner, Yenni...28–32

Z

Zero-Velocity-Error Servomechanisms, Two Types of Levinson....19–27

2. Author Index

Alger, P. L. Disc....122
Apple, J. M.; E. W. Lusby. Shipboard Use of 400-Cycle Electric Power. (56–184)..116–21; disc. 124
Arnold, O. B. Disc....42
Aroyan, George F.; K. G. Lakey. Distributing Large Single-Phase Loads in Limited Capacity 3-Phase Power Systems. (56–625)....249–52

B

Backenstos, H. B. Economic Studies Applied to Industrial Power Systems. (56–22)....61–7; disc. 69
Barlow, Erle P. Disc....188, 197, 223
Bass, Robert W. Disc....231
Benner, R. H.; T. B. Jones. Influence of Atmospheric Water Vapor on High-Current D-C Arcs. (56–109)....162–6
Bertram, John E. Factors in the Design of Digital Controllers for Sampled-Data Feedback Systems. (56–209)....151–8
Bharucha, B. H.; F. W. Sentelka. Disc....411
Biernson, George A. A General Technique for Approximating Transient Response from Frequency Response Asymptotes. (56–640)....253–73
Bills, G. W. Disc....389
Boggess, P. F.; R. D. Jessee. Negative-Sequence Voltage Detection on Aircraft A-C Electric Systems. (56–440)....171–5
Boyer, J. L.; A. P. Colclasure. Sealed Ignitron Principles Extended to Larger Tubes. (56–173)....125–8; disc. 129
Brereton, D. S.; E. P. Pesbody. Electric System for a Nuclear Reactor. (56–674)....296–306
Brown, H. F. Disc....443
Bryant, C. M.; H. S. Sacchini, M. Rosenblatt. Recent Trends in Aircraft Ground Power Electric Equipment. (56–445)....214–23; disc. 223
Buchanan, R. A.; W. R. Hodgson. Rectifier Power Supply for a Modern Rod Mill. (56–44)....70–4; disc. 74

C

Caldwell, S. C.; R. T. Smith. Disc....180, 187
Chao, Stanley K. Design of a Contactor Servo Using Describing Function Theory. (56–456)....223–31; disc. 233
Clegg, John C. A Time-Dependent Nonlinear Compensating Network. (56–778)....306–08; disc. 308
Colclasure, A. P.; J. L. Boyer. Sealed Ignitron Principles Extended to Larger Tubes. (56–173)....125–8; disc. 129
Convery, J. F. Disc....124
Courcoulas, J. H.; J. M. Ham. Incremental Control Equations for Tandem Rolling Mills. (56–656)....363–70; disc. 371
Cox, H. N.; L. G. Levoy. Electric Power Systems for Steel Plants. (56–657)....339–48; disc. 349
Creveling, Cyrus J. Disc....213
Curtis, L. B. Disc....441

D

Dalasta, D. Disc....69
Dallas, J. P.; L. B. Kilman. High Potential Testing of Aircraft Equipment Electric Insulation. (56–441)....189–93
Dalziel, C. F.; F. P. Massoglia. Let-Go Currents and Voltages. (56–111)....49–55; disc. 55
Dawson, C. H. Disc....96
Denlston, W. L. Disc....204
Derr, W. A.; W. L. Metz. Centralized Control of Power Systems in the Steel Industry. (56–731)....332–7; disc. 338
Dickinson, W. H. Disc....68
Dister, H. A., Jr. Important Factors Influencing the Choice Between a Synchro and a Potentiometer as an Angular Position Pickoff. (56–449)....198–204; disc. 206
Donaldson, Lieut. C. E., III. Disc....123
Dortort, I. K. Harmonic Analysis by Direct Area Measurements. (56–188)....16–19
Dortort, I. K. Disc....47

E

Eckel, James R., Jr. Disc....419
Evans, W. G.; W. G. Hall, R. I. Van Nice. Magnetic Logic Circuits for Industrial Control Systems. (56–91)....166–71
Exner, D. W. Disc....193

F

Farmanfarma, G. Disc....149
Ferguson, R. W.; W. A. Munson. Disc....188
Findley, J. D.; L. D. McConnell. A New High-Power Cathode Circuit Breaker. (56–185)..39–42; disc. 43
Flegge-Lotz, Lu. Disc....289
Fraser, H. J.; E. G. Hopkins. An Automatic Output Regulator for Power Oscillators. (56–15)....36–8
Frick, C. W. Disc....99
Friedland, Bernard. A Technique for the Analysis of Time-Varying Sampled-Data Systems. (56–962)....407–11; disc. 413
Frosztega, H. G. Disc....338
Fuchs, Abraham M. Disc....139, 248

G

Gallagher, J. J. An Air-Cleaning Blower for Ventilating Traction Equipment. (56–105)....159–61
Gatewood, Arthur R. Disc....121
Grafler, Harold. Disc....205

H

Hague, C. S.; R. V. Wachter, C. R. Marcum. Test and Operating Experience with the Triple-Diametric Rectifier. (56–186)....100–04; disc. 104
Hall, William A. Disc....123
Hall, W. G.; W. G. Evans, R. I. Van Nice. Magnetic Logic Circuits for Industrial Control Systems. (56–91)....166–71
Ham, James M. Disc....370
Ham, J. M.; J. H. Courcoulas. Incremental Control Equations for Tandem Rolling Mills. (56–656)....363–70; disc. 371
Hatch, P. H. Modernization of the Long Island Rail Road Passenger-Car Fleet. (56–979)....395–401
Hayes, Arnold E. Disc....205
Headley, J. F. Disc....337
Herskind, C. C. Disc....105, 129
Hesenberg, W. C. F.; W. N. Jenkins. Disc....371
Heumann, G. W. Applying Industrial Controllers for Safe Operation. (56–269)....291–5
Higgins, Thomas J. Disc....289, 405
Hodgson, W. R.; R. A. Buchanan. Rectifier Power Supply for a Modern Rod Mill. (56–44)....70–4; disc. 74
Hopkin, A. M.; M. Iwama. A Study of a Predictor-Type Air-Frame Controller Designed by Phase-Space Analysis. (56–198)....1–9
Hopkin, A. M.; K. Ogata. Disc....233
Hopkins, E. G.; H. J. Fraser. An Automatic Output Regulator for Power Oscillators. (56–15)....36–8

I

Inglis, J. G. Rapid Transit Progress in Toronto. (56–762)....352–4
Iwama, Morimi; A. M. Hopkin. A Study of a Predictor-Type Air-Frame Controller Designed by Phase-Space Analysis. (56–198)....1–9

J

Jenkins, W. N.; W. C. F. Hesenberg. Disc....371
Jessee, R. D.; P. F. Boggess. Negative-Sequence Voltage Detection on Aircraft A-C Electric Systems. (56–440)....171–5
Jeynes, Paul H. Disc....67
Jones, T. B.; R. H. Benner. Influence of Atmospheric Water Vapor on High-Current D-C Arcs. (56–109)....162–6
Jury, Eliahu I. Hidden Oscillations in Sampled-Data Control Systems. (56–790)....391–5
Jury, Eliahu I. Synthesis and Critical Study of Sampled-Data Control Systems. (56–208)..141–9; disc. 151
Jury, Eliahu I.; W. Schroeder. Discrete Compensation of Sampled-Data and Continuous Control Systems. (56–644)....317–25
Jury, E. I.; W. Schroeder. Disc....115, 158

K

Kahle, H. A. Parallel Operation of Aircraft A-C Generators: Reactive Load Division and Selective Protection. (56–435)....175–80; disc. 181
Kahle, H. A.; H. M. McConnell. Analogue Computer Methods Applied to Steady State A-C System Problems. (56–766)....279–82
Kazda, Louis F.; R. E. Kuba. A Phase-Space Method for the Synthesis of Nonlinear Servomechanisms. (56–455)....282–9; disc. 290
Kilman, L. B.; J. P. Dallas. High Potential Testing of Aircraft Equipment Electric Insulation. (56–441)....189–93
Kimball, R. L. Disc....442
King, E. B. Disc....188
Koenig, W. P. Disc....186
Kouwenhoven, W. B. Disc....55
Krasin, K. A. Disc....55
Kroneberg, A. A. Voltage Dips and Flicker. (56–837)....349–51; disc. 352
Ku, Y. H. Analysis of Servomechanisms with Nonlinear Feedback Control. (56–954)..402–05; disc. 495
Kuba, Richard E.; L. F. Kazda. A Phase-Space Method for the Synthesis of Nonlinear Servomechanisms. (56–455)....282–9; disc. 290

L

Lakey, Keith G.; G. F. Aroyan. Distributing Large Single-Phase Loads in Limited Capacity 3-Phase Power Systems. (56–625)....249–52
Lawrence, Robert F. Disc....352
Lawton, William E.; H. Winograd. Methods of Measuring Arc-Drop Voltage on Mercury-Arc Rectifiers. (56–187)....43–7; disc. 48
Levinson, Emanuel. Two Types of Zero-Velocity-Error Servomechanisms. (56–195)....19–27
Levoy, L. G. Disc....351
Levoy, L. G.; H. N. Cox. Electric Power Systems for Steel Plants. (56–657)....339–48; disc. 349
Lister, C. A. Interruption Tests on High-Voltage Air Break Contactor. (56–31)....56–9; disc. 60
Lloyd, Wayne B. Disc....83
Ludwig, H. C. Cathode Instability in Argon Atmospheres. (56–110)....45–5
Lusby, E. W.; J. M. Apple. Shipboard Use of 400-Cycle Electric Power. (56–184)..116–21; disc. 124

M

Maitra, K. K.; P. E. Sarachik. Digital Compensation of Continuous-Data Feedback Control Systems. (56–210)....107–14; disc. 115
Marcum, C. R.; R. V. Wachter, C. S. Hague. Test and Operating Experience with the Triple-Diametric Rectifier. (56–186)....100–04; disc. 104
Markowitz, O. Development of Definitions for Voltage Modulation Characteristics in 380- to 420-Cycle-Per-Second Aircraft Electric Systems. (56–434)....193–7; disc. 197
Massoglia, F. P.; C. F. Dalziel. Let-Go Currents and Voltages. (56–111)....49–55; disc. 55
Maxwell, H. S. Disc....371
Mayer, V. W. Disc....121
McConnell, H. M. Disc....185
McConnell, H. M.; H. A. Kahle. Analogue Computer Applied to Steady State A-C System Problems. (56–766)....279–82
McConnell, L. D.; J. D. Findley. A New High-Power Cathode Circuit Breaker. (56–185)..39–42; disc. 43
McDonald, D. C. Disc....290
McFarland, G. L. Disc....60
McGee, P. A. Disc....442
Metz, W. L.; W. A. Derr. Centralized Control of Power Systems in the Steel Industry. (56–731)....332–7; disc. 338
Miller, W. E. Disc....74
Mishkin, E. Feedback Techniques Applied to the Single-Phase Transformer. (56–791)....374–6; disc. 376
Morton, L. W. Disc....128
Morton, L. W.; A. Schmidt, Jr. Disc....104
Mullin, F. J. Disc....308
Munson, W. A.; R. W. Ferguson. Disc....188
Myers, E. H. Disc....351
Myles, A. H. A 2-Motor A-C Mine Hoist Control System. (56–84)....10–16

N

Nelswander, R. S. An Experimental Treatment of Nonlinear Servomechanisms. (56–777)....308–16
Nikiforuk, P. N. Response of a Particular Nonlinear Control System to Random Signals. (56–924)..419–22

Nikiforuk, P. N.; J. C. West. The Frequency Response of a Servomechanism Designed for Optimum Transient Response. (56–452)....234–9
Nonoto, Akira. Disc....96

O

Ogata, K.; A. M. Hopkin. Disc....233
Oman, Henry. Frequency Modulation and Load-Division Instability in 400-Cycle Aircraft Electric Systems. (56–436)....181–5; disc. 188

P

Panser, M. Envelope Transfer Function Analysis in A-C Servo Systems. (56–792)....274–9
Partlow, J. G. Disc....122
Peabody, E. P.; D. S. Brereton. Electric System for a Nuclear Reactor. (56–674)....296–306
Phillips, R. A. Analysis of Tandem Cold Reduction Mill with Automatic Gauge Control. (56–638)....355–63; disc. 370
Price, J. C. Power Supply for Commercial-Frequency Railroad Electrification. (56–977). 432–41; disc. 444

Q

Quick, C. E. Disc....351

R

Robinson, D. B. Disc....55
Rosenblatt, M.; H. S. Sechrist, C. M. Bryant. Recent Trends in Aircraft Ground Power Electric Equipment. (56–445)....214–23; disc. 223

S

Salzer, John M. Disc....413
Sarachik, P. E.; K. K. Maitra. Digital Compensation of Continuous-Data Feedback Control Systems. (56–210)....107–14; disc. 115
Satyendra, K. N. Describing Functions Representing the Effect of Inertia, Backlash, and Coulomb Friction on the Stability of an Automatic Control System. Part I. (56–453)....243–8
Saunders, R. M. Disc....370, 371
Schaefer, L. P. A Pressure Control Switch with a Pulsation Filter. (56–932)....426–32
Schieber, L. Variably Damped Servomechanisms. (56–958)....414–18
Schmidt, A., Jr.; L. W. Morton. Disc....104
Schroeder, William; E. I. Jury. Discrete Compensation of Sampled-Data and Continuous Control Systems. (56–644)....317–25
Schroeder, W.; E. I. Jury. Disc....115, 158
Sechrist, H. S.; M. Rosenblatt, C. M. Bryant. Recent Trends in Aircraft Ground Power Electric Equipment. (56–445)....214–23; disc. 223
Semelka, F. W.; B. H. Bharucha. Disc....411
Shearer, J. L. Disc....83
Shutter, J. D. Disc....121
Silberberg, M. Y. A Note on the Describing Function of an Element with Coulomb, Static, and Viscous Friction. (56–960)....423–5
Skinner, G. M.; D. M. Yenni. Inert Gas-Shielded Welding Arc Behavior and Metal Transfer Characteristics. (55–112)....28–32
Sklansky, Jack. Disc....150, 413
Siemon, Gordon R. Disc....419
Smith, Edward P. Transient Characteristics of D-C Metal-Rolling Motors and Generators. (56–639)....326–31; disc. 331
Smith, R. T.; S. C. Caldwell. Disc....180, 187
Stallard, David V. Analysis and Performance of a Valve-Controlled Hydraulic Servomechanism. (56–197)....75–83; disc. 84
Stone, K. A. Disc....83
Stout, T. M. A Step-by-Step Method for Transient Analysis of Feedback Systems with One Nonlinear Element. (56–779)....378–89; disc. 389
Stout, T. M. Block Diagram Transformations for Systems with One Nonlinear Element. (56–196)....130–9; disc. 139
Stout, T. M. Disc....376

T

Tull, I. N. Disc....348

V

Vallese, L. M. Comparison of Backlash and Hysteresis Effects in Second-Order Feedback Systems. (56–454)....240–3
Van Nice, R. I.; W. G. Evans, W. G. Hall. Magnetic Logic Circuits for Industrial Control Systems. (56–91)....166–71

W

Wachter, R. V.; C. S. Hague, C. R. Marcum. Test and Operating Experience with the Triple-Diametric Rectifier. (56–186)....100–04; disc. 104
Welch, A. U. Disc....55
West, J. C.; P. N. Nikiforuk. The Frequency Response of a Servomechanism Designed for Optimum Transient Response. (56–452)....234–9
Wilson, B. J. Analyzing Missile Electric System Reliability. (56–450)....206–12; disc. 213
Winograd, H. Disc....106
Winograd, Harold; W. E. Lawton. Methods of Measuring Arc-Drop Voltage on Mercury-Arc Rectifiers. (56–187)....43–7; disc. 48
Wright, F. H.; S. A. Zimmermann. Evaluation of Radio Influence Voltages in Fluorescent Lighting Installations. (56–75)....96–9; disc. 99

Y

Yeh, Victor C. M. Synthesis of Feedback Control Systems by Gain-Contour and Root-Contour Methods. (56–211)....85–96
Yenni, D. M.; G. M. Skinner. Inert Gas-Shielded Welding Arc Behavior and Metal Transfer Characteristics. (55–112)....28–32

Z

Zadeh, L. A. Disc....139
Zimmermann, S. A.; F. H. Wright. Evaluation of Radio Influence Voltages in Fluorescent Lighting Installations. (56–75)....96–9; disc. 99

A Study of A Predictor-Type Air-Frame Controller Designed by Phase-Space Analysis

A. M. HOPKIN
MEMBER AIEE

MORIMI IWAMA
NONMEMBER AIEE

IN THE past, various forms of optimum relay servomechanisms have been proposed.[1-6] The connotation of the word optimum is given by the definition: The optimum performance of a servomechanism will be that behavior in which the system returns to rest with zero or minimum error in the shortest time following a step input.[1]

In general, these devices apply the forcing function by means of a relay or narrow-proportional band-saturating amplifier in such a way that the system spends most of its transient operating time with the forcing function at one extreme or the other. The application of the forcing function is controlled by a predictor device which utilizes error and various order derivative signals to determine the time at which the forcing function should go from one polarity maximum to the other, or to zero, or in some cases to a linear mode of operation about zero error.

The predictor usually uses an analogue representation of phase-plane or phase-space trajectories to determine switching time. In general a certain trajectory or set of trajectories, or modifications thereof, determine critical boundaries. Switching should take place when the locus representing the system operating condition crosses a critical boundary for optimum system performance.

For systems whose characteristic equation is of second order, the phase plane provides a fairly simple solution, even in the nonlinear case. For third- and higher order systems, the phase-space solution becomes increasingly complex. Even for the general third-order case, the analysis and design of a predictor can present difficulties.

Nomenclature

a_1, a_2, a_4 = constant of integration
a_x = pitch linear acceleration, feet per second^2
a_{x_c} = pitch linear acceleration command
$A_x(s)$ = Laplace transform of a_x
$A_{x_c} = a_{x_c}$ (the machine value)
$b(t)$ = actual forcing function
$B(T)$ = forcing function (the machine value)
e or $e(t)$ = system error of a servomechanism
$E = e$ (the machine value)
$\dot{e}$ = time derivative of the system error de/dt
$\ddot{e}$ = second time derivative of the system error d^2e/dt^2
$F_c(t)$ = forcing function generated by the predictor
F_m = maximum magnitude of $Fe(t)$
$g(s)$ = system transfer function
$F_c(s)$ = Laplace transform of $Fe(t)$
$g_1(s)$ = air-frame transfer function
K = system forward gain
K_v = valve actuator mechanism gain
K_1, K_{1b}, K_2, K_{2b} = system parameters
$\mathcal{L}(\ )$ = Laplace transform of the term in brackets
$M_{\delta E}, M_w, M_q$ = aerodynamic coefficient; see Table I
$M_1, M_3, M_2, M_4, M_{1b}, M_{2b}, M_{3b}, M_{4b}$ = conversion factors
q = pitch angular velocity, radians per second
$\dot{q}$ = pitch angular acceleration dq/dt, radians per second^2
$Q(S) = q$ (Laplace transform)
s = Laplace transform operator
t = time
T = machine time
u = forward velocity along X axis, feet per second
w = velocity along Z axis, feet per second
$\dot{w}$ = acceleration along Z axis dw/dt, feet per second^2
$W(S) = w$ (Laplace transform)
$Z_{\delta E}, Z_w$ = aerodynamic coefficient; see Table I
δE = air-frame deflection, radians
$\Delta E(s) = \delta E$ (the Laplace transform)
$\Delta E(T)$ or $\Delta E = \delta E$ (the machine value)
$\omega_n, \omega, \omega_0$ = natural frequency of the system
ζ = damping ratio
ϵ = Naperian base (2.71828. .)
α = damping coefficient
β = damped frequency of the system
θ = constant of integration (angle, radians)

The System Controlled

The design of a predictor device for the control of the pitch linear acceleration of a tail-controlled air frame in response to a pitch-acceleration command is considered. The response of the air frame is inherently highly underdamped and conventional linear control and compensation result in slowly dying long-period oscillations. Fig. 1 shows the co-ordinate system for air-frame equations.

Fig. 2 shows the block diagram of the pitch-acceleration control system. It consists of the air frame, the elevator member, the hydraulic actuator, and a solenoid-operated control valve.

Appendix I gives the air-frame differential equations and a description of the valve and hydraulic actuator characteristics. Appendix II shows the derivation of the air-frame transfer function. Since the air-frame natural frequency is of the order of 0.3 to 1 radians per second, and the actuator natural frequency and the valve lag correspond to frequencies in the order of 20 radians per second, the valve and actuator transfer function may be approximated as being that of a simple integrator. As a result, the system open-loop transfer function may be simplified to the form

$$g(s) = \frac{A_x(s)}{F_c(s)} =$$

$$K_v \frac{s^2 Z_{\delta E} - s M_q Z_{\delta E} + s(Z_w M_{\delta E} - M_w Z_{\delta E})}{s(s^2 + 2\zeta\omega_n s + \omega_n^2)} \quad (1)$$

The actual system open-loop transfer function is of sixth order, but the third-order approximation is assumed to be a reasonably accurate representation of the system. Fig. 3 shows the block diagram of the system based on the third-order approximation. The air-frame parameters are a function of flight conditions. These vary over the range indicated in Table I.

The Phase-Space Description of the Open-Loop System

The dynamic behavior of the control system can be described by all of the possible space trajectories representing the responses to a step-forcing function in a 3-dimensional phase space. The co-ordinates along the three axes may be system output acceleration, first derivative of acceleration, and second derivative of acceleration. However, by appropriate transformation, the co-ordinates may also consist of combinations of these three variables. Appendix II contains the derivation of the system differential equation for the system error following a step input forcing function

Paper 56-198, recommended by the AIEE Feedback Control Systems Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Winter General Meeting, New York, N. Y., January 30–February 3, 1956. Manuscript submitted October 17, 1955; made available for printing December 27, 1955.

A. M. HOPKIN and MORIMI IWAMA are with the University of California, Berkeley, Calif.

The work reported here was supported by the Boeing Airplane Company, Seattle, Wash. Typical air-frame data and other information and assistance were supplied by R. W. Ullman, G. H. Hage, and R. Von Tobel, all of Boeing Airplane Company.

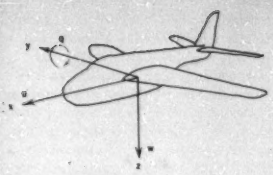


Fig. 1. Co-ordinate system for air-frame equations

and its solution. The system-error differential equation is

$$\ddot{e}+2\zeta\omega_n\dot{e}+\omega_n{}^2e=\omega_n{}^2a_{z_c}\mp K_1K_2F_m-K_2[\delta E(0)\pm K_2F_mt] \quad (2)$$

where

$$K_1=-M_qZ_{\delta E} \quad (3)$$

$$K_2=u(Z_wM_{\delta E}-M_wZ_{\delta E}) \quad (4)$$

F_m represents positive or negative maximum value of the forcing function (either extreme of displacement of the 4-way control valve driver). Equation 2 can be rewritten in the form

$$\ddot{b}(t)=\ddot{e}+2\zeta\omega_n\dot{e}+\omega_n{}^2e=\omega_n{}^2a_{z_c}\mp K_1K_2F_m-K_2[\delta E(0)\pm K_2F_mt] \quad (2A)$$

Equation 2(A) defines a new quantity $b(t)$ which is a linear combination of e, $\dot{e}$, and $\ddot{e}$. Also $b(t)$ can be expressed as a linear combination of the instantaneous input command a_{z_c}, the instantaneous control surface deflection $\delta E=\delta E(0)\pm K_2F_mt$, and a constant. Therefore $b(t)$ can readily be obtained as a voltage signal from information representing the input command signal and the control surface deflection, while it is difficult to determine accurately a voltage signal representing $\ddot{e}$. For this reason the phase space is set up in terms of e, $\dot{e}$ and $b(t)$ rather than in terms of e, $\dot{e}$, and $\ddot{e}$. Trajectories in the new phase space involving $b(t)$ represent the same quantities as in the original phase space, but they are warped as a result of the linear transformation involved.

In the original phase space the switching criteria for optimum behavior is that which was determined for a third-order system by Bogner and Kazda.[2] The last reversal of the switching device must occur when the phase-space trajectory representing the system dynamic state intersects the particular reversed trajectory which goes through the origin (the critical switching boundary line). The system trajectory ideally will then follow the critical switching boundary line to the phase-space origin, which represents the system state of zero error.

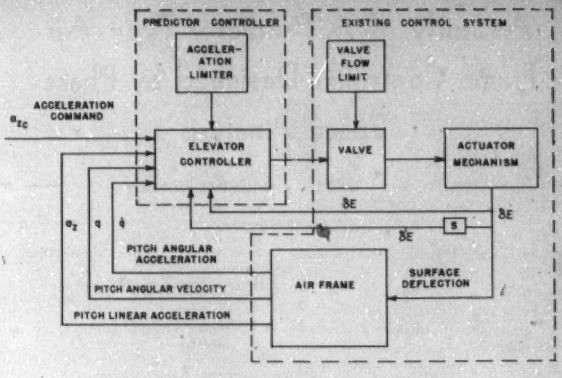


Fig. 2. Block diagram of pitch acceleration control system

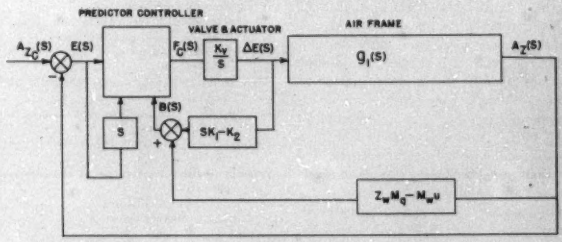


Fig. 3. Block diagram of pitch acceleration predictor control system using a third-order approximation

To guarantee that the system trajectories will intersect the critical boundary line, it is necessary to have a first reversal which occurs when the system trajectory intersects the critical switching surface made up of all the possible trajectories (resulting from the first reversal) which will intersect the critical switching boundary line.

For the case in this paper where the third phase-space co-ordinate $\ddot{e}$ is replaced by $b(t)$, the procedure given in the foregoing holds, but the phase-space trajectories must be determined in the new co-ordinate system.

In terms of these co-ordinates, representative phase-space trajectories were computed for both positive and negative maximum values of the forcing function, with the use of equation 21 of Appendix II together with the appropriate constants. Typical trajectories so obtained are shown in Fig. 4. For evaluation of the critical switching boundaries, the particular trajectories which pass through the points $K_1K_2F_m$ units away from the origin along the $b(t)$ axis are of special interest. If the forcing function $F_c(t)$ is made zero at this point, then the error of the system remains zero until the command signal A_{z_c} changes. In subsequent discussion, the curves A_1 and A_2 of Fig. 5 will be assumed to be such critical switching trajectories for the positive and negative maximum forcing function F. These critical switching trajectories determine the time of the last of the two reversals of $F_c(t)$ required for optimum system response to a step command input A_{z_c}. The critical switching surfaces are composed of the particular trajectories which terminate on the curves A_1 and A_2. These switching surfaces form the boundaries for the positive and negative output regions of the predictor controller and determine the time of the first reversal of $F_c(t)$.

The switching sequences for optimum operation are described in the phase-space as follows: Assume that the point P_1 in Fig. 5 describes the system conditions in the phase space, and that P_1 lies in the

Table I. Typical Values of Aerodynamic Coefficients

Flight Condition Code	$-M_{\dot{\alpha}}$, Radians per Second[3]	$-M_{\Delta E}$, Radians per Second[2]	$-M_w$, Radians per Second[3]	Forward Velocity u,	$-Z_{\Delta E}$, Feet per Second[2]	$-Z_{\dot{\alpha}}$, Feet per Second[2]	
	Radians per Second	Feet per Second	Feet per Second	Feet per Second	Radians	Feet per Second	Dynamic Pressure
21.4	0.1095	0.632	0.00390	145.0	5.34	0.244	
22.0	0.1137	0.810	0.00363	207.0	5.92	0.274	
22.6	0.1110	1.005	0.00350	269.0	7.14	0.298	Maximum
41.4	0.0473	0.258	0.00158	135.8	2.15	0.1054	
42.0	0.0491	0.3265	0.00157	194.0	2.39	0.178	
42.6	0.0480	0.405	0.00151	252.0	2.875	0.129	
61.4	0.0183	0.0981	0.000646	135.8	0.827	0.0407	Minimum
62.0	0.0189	0.1256	0.000610	194.0	0.920	0.0454	
62.6	0.0185	0.1560	0.000580	252.0	1.10	0.0493	

negative F_m region. The predictor controller will provide a signal to drive the control valve to the negative maximum flow condition. Consequently, the response trajectory will follow the negative F_m curve until the critical switching surface is intersected at point P_2 of Fig. 5. At this time the predictor controller will provide the signal to drive the control valve to the positive maximum flow condition, and the system trajectory will follow the curve B_2 corresponding to the positive F_m condition until the curve A_2 is reached. The predictor will again put out a negative F_m signal, driving the control valve to a negative maximum flow condition, and the system trajectory will follow the A_2 curve of Fig. 5 until the trajectory intersects the $b(t)$ axis at $-K_1K_v F_m$ units from the origin. At this point the predictor controller must provide the signal $F_c(t)=0$ to bring the control valve to the zero flow condition.

The Analogue Method of Finding the Critical Switching Boundaries

The method of calculating the phase-space trajectories from analytical expressions for e, $\dot{e}$, and $b(t)$ is exceedingly tedious. Therefore most of the trajectories were obtained by the technique of "reversing the time axis" with the use of an analogue computer. The technique volves writing the system differential equation and replacing t with $(-t)$, where it appears, to give an expression of the form

$$\ddot{e}-2\zeta\omega_n\dot{e}+\omega_n e=b(-t)$$

By setting up the analogue of this equation on the computor and using as the initial condition the final desired steady-state condition $e(0)=0$, $\dot{e}(0)=0$, $b(0)=\mp F_m K_v K_1$, and applying $F_c(t)=\pm F_m$, the trajectories A_1 and A_2 representing the critical boundaries of second reversal are traced by the computer in a reversed direction. By switching $F_c(t)$ to the negative of its original condition at successively later times on repeated runs, the various curves representing the critical switching boundary surfaces for the first reversal are set up. The machine variable values obtained from the analogue-computor study differed from the actual values of the variables by linear scale factors. The machine variable values are used in all of the figures, i.e., T, $B(T)$, and $E(T)$ correspond respectively to t, $b(t)/K_2$, and $e(t)$ of the actual system equations. ($b(t)/K_2$ rather than $b(t)$ was used to secure 1-to-1 correspondence with the air-frame deflection δE or ΔE.)

Fig. 6 shows a representation of the optimum switching boundaries, and Fig. 7 shows a 3-dimensional wire model made up to aid in visualizing the phase-space trajectories. It is to be noted that the method of reversing the time axes is not limited to linear systems, but can be used for any system which can be accurately represented on an analogue computor.

Setting Up the Predictor Control Device

The practical predictor control device must be some type of analogue of the phase space which is set up to recognize when the system trajectory intersects the critical boundary, and then sends out the necessary switching commands. The devising of such an analogue is complicated in the case at hand by the fact that the space is 3-dimensional and the air-frame transfer function is oscillatory and highly underdamped, with the result that many of the trajectories are double valued. The nature of the system is such that the shape of the projections of the switching surface change as the distance along the $B(T)$ axis changes, as can be seen from Fig. 6. For these reasons the use of biased diodes and similar nonlinear function-generating devices was considered impractical if not impossible. The function generator might have been of the form of a cam surface with 2 degrees of freedom. However, this method of approach was not used because of the difficulty of later changing the cam contour.

For this study made on an analogue computor, a workable solution was found in the form of a cathode-ray oscillograph-photocell combination[1] provided with a mask which changed shape as $B(T)$ varied to simulate the effect of the varying shape of the projections of the critical boundary surfaces with $B(T)$. The mask consisted of three properly shaped moving plates linked to a shaft, the rotation of

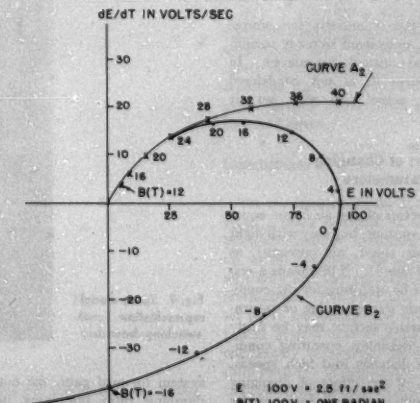


Fig. 4. Analytical results for curves A_2 and B_2 for flight condition A

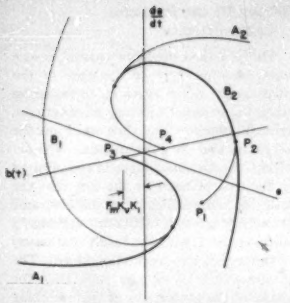


Fig. 5. Phase-space co-ordinate system for switching boundary study with typical trajectories shown

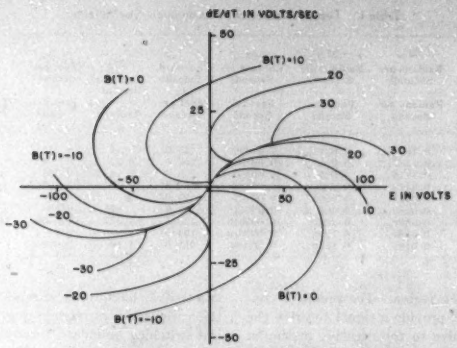


Fig. 6. Optimum switching boundaries for flight condition A

which was linearly proportional to $B(T)$. For any given value of $B(T)$, the plates overlap to form a moderately accurate representation of the corresponding switching surface projection. The inputs to the X and Y axis of the oscilloscope are acceleration-error voltage and acceleration-error derivative voltage respectively. For negative values of $B(T)$, the photocell generated $+F_m$ when the spot was behind the mask and $-F_m$ when the spot was visible. Figs. 8 and 9 show the actual device.

In application, the masks are shaped to generate the boundary for negative values of $B(T)$. For positive values of $B(T)$, a relay operates to effect a skew-symmetrical pattern by reversing the polarities of all of the inputs in the function generator.

This generator is massive but simple to set up and lends itself to ready modification of the critical boundaries. In its present form it is not considered practical for actual use with an air frame.

The Problem of Changing System Parameters

The parameters of the air-frame equations are not constant, but vary with flight conditions (air speed, air pressure), as indicated in Table II. This means a system set up to be optimum for the conditions of the air frame leaving or returning to the earth's surface may be fairly far from the optimum operating condition at high altitudes and high speeds. Investigation of the effect of changing flight conditions on the air frame under consideration showed that the primary effect was to change the natural frequency of the air-frame response and the system forward gain, the corresponding damping ratio remaining almost constant; see Table II and Fig. 10 for an idea of how these parameters vary with flight conditions.

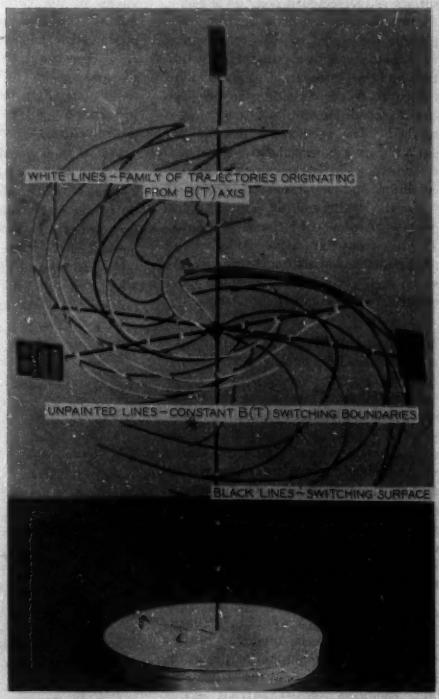

WHITE LINES – FAMILY OF TRAJECTORIES ORIGINATING FROM B(T) AXIS
UNPAINTED LINES – CONSTANT B(T) SWITCHING BOUNDARIES
BLACK LINES – SWITCHING SURFACE

Fig. 7. Space model representation of switching boundary

Appendix III shows that, if the damping factor can be considered constant, the effect of varying natural frequency and the forward gain can be compensated for by a linear change of gain in the predictor in-

Fig. 8 (above). Front view of function generator

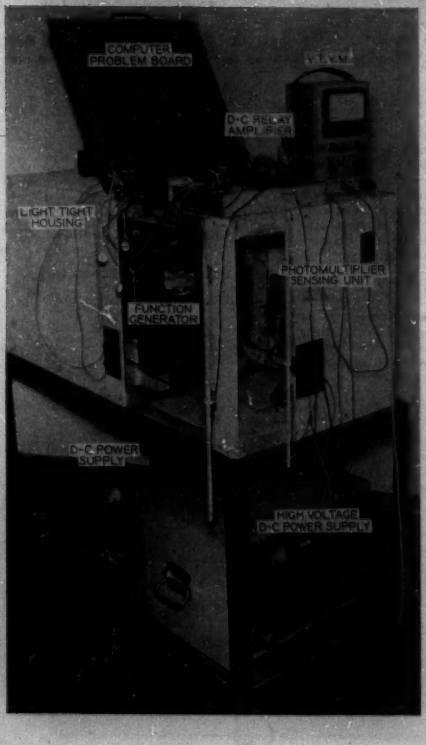


Fig. 9 (left). Predictor controller, assembled

put channels without making other changes in the phase-space analogue component. Automatic compensation for changing flight conditions would then require a device which recognized the change of natural frequency associated with the change of flight conditions, and which readjusted the gains of the inputs to the phase-space analogue in the controller unit accordingly.

Results

A predictor controller was designed for an air frame operating with the coefficients of flight condition A tabulated in Table III. It was set up to machine scale to operate with the analogue computor simulating the air frame, valve, and actuator dynamical properties. The predictor switching boundaries were only approximate, and the mask-positioning device had some dynamical error in following the $B(T)$ signal. However, as shown in Fig. 11, the comparison of the ideal predictor system step input response with that of the system set up was reasonably close.

Fig. 12 shows the effect of changing air-frame damping ratio on the response of the analogue system with the predictor unit designed for a damping ratio ζ of 0.1. It will be noted that a variation of ζ from 0.1 to 0.3 does not affect the system response appreciably. Table II and Fig. 10 show that for the system under consideration the damping ratio varies even over the smaller range. Thus, the assumption that the effect of change of damping ratio upon the performance is negligible is valid.

It will be noted from Fig. 11 that, for most initial conditions and input step amplitudes normally expected, the predictor control system will bring the error to approximately zero after one overshoot swing. Because the air frame is highly oscillatory, the actual number of overshoot swings will depend on the initial conditions at the time of application of the step and the step amplitude.

Fig. 11 also shows that the predictor controller relay chatters when the error approaches zero. It is assumed that better performance could be obtained by replacing the relay with a device which would give a linear mode of operation for small errors.

Conclusions

The predictor controller device designed gives a good response as an air-frame pitch-acceleration controller when set up to operate with an analogue of an air-frame and control-surface-positioning mechanism. The particular predictor is set up to handle the case where the controlled system was of third order with one highly oscillatory mode (damping ratio $\zeta = 0.1$). Hence, it requires a function generator which will generate a function of two variables $E(T)$ and $\dot{E}(T)$, and change the form of this function in a prescribed fashion as a third variable $B(T)$ changes. The unit constructed

Table II. System Parameters for Typical Flight Conditions

Flight Condition Code	ω_n, Radians / Seconds	ζ, Dimensionless	K_1, Feet / Seconds^2	K_2, Feet / Seconds^4	Dynamic Pressure
21.4	0.772	0.230	−0.585	19.3	
22.0	0.885	0.219	−0.673	41.4	
22.6	0.987	0.207	−0.793	73.4	Maximum
41.4	0.484	0.158	−0.102	3.18	
42.0	0.560	0.203	−0.117	10.56	
42.6	0.625	0.142	−0.138	12.07	
61.4	0.298	0.099	−0.0151	0.468	Minimum
62.0	0.345	0.0933	−0.0175	1.105	
62.6	0.384	0.0883	−0.0204	1.773	

satisfies the requirements reasonably well but, in its present form, is bulky and is considered to be impractical for use on an actual air frame. Further work must be done to produce a practical form of function generator for this type of predictor control system.

It was found that, for the system under consideration, a linear change of scale factor (gain) of the inputs to the predictor unit is sufficient to compensate for changes of air-frame natural frequency and gain caused by changing flight conditions. The change of damping ratio under these circumstances produced negligible effects on the predictor performance, primarily because this change was relatively small. Thus a predictor control system of the optimum relay type can be used to control the linear pitch acceleration of an air frame and, with appropriate inputs indicating the effect of changing flight conditions, can be automatically adjusted to compensate for parameter changes resulting from these changing flight conditions.

Appendix I. Differential Equations and Valve and Hydraulic Actuator Characteristics

Air-Frame Equations for Z Components

Linearized approximate equations are used in this study. It is assumed that control surface forces are perpendicular to the line of flight and proportional to the control surface deflection and that there is no interaction between control effects on different axes. The resulting constants or, aerodynamic coefficients, then are truly constant only for small control surface deflections and for only one set of flight conditions. Typical Aerodynamic coefficients in generalized form for varying flight conditions are given in Table I. All data were supplied in reduced generalized form by the Boeing Airplane Company.

The linearized equations for an air frame of one unit of mass and one unit of rotary inertia are

$$Z_{\delta E}\delta E = -Z_w w + \dot{w} - uq \qquad \text{(summation of forces)} \quad (5)$$

$$M_{\delta E}\delta E = -M_w w - M_q q + \dot{q} \qquad \text{(summation of moments)} \quad (6)$$

$$a_Z = \dot{w} - uq \qquad \text{(effect of axes moving with air frame)} \quad (7)$$

a_Z is the acceleration normal to the longitudinal axis of the air frame. a_Z is related to $\dot{w}$, the acceleration along Z axis of the moving co-ordinate system, shown in Fig. 1 and equation 10. The term uq in equation 10 is the result of an instantaneous rotation of the co-ordinate system.

Hydraulic Valve

The valve may be represented as a simple lag with a relatively small time constant in the order of 0.05 second. The flow is a linear function of the error signal, except that flow is limited such that the equivalent maximum surface velocity is in the order of 0.04 radian per second. However, the flow limit should be considered as a parameter in the elevator controller design.

Actuator Mechanism

The transfer function of the actuator mechanism may be approximated as being an integrator plus a second-order system having a natural frequency of 3.5 cycles per second and a damping ratio of 0.05.

Limits on Air-Frame Motion

The predictor servo must be capable of controlling the air-frame motion so as not to exceed certain structural limitations. These structural limits may be expressed in terms of maximum allowable pitch linear acceleration $a_{Z\max}$, and pitch angular acceleration $\dot{q}_{\max}$. These limits should be treated as parameters. Typical values are

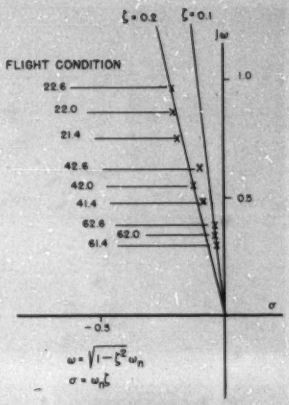


Fig. 10. System parameters for typical flight conditions shown by means of S-plane root locations

$z_{\max} = 2.25$ feet per second^2

$\dot{q}_{\max} = 0.05$ radian per second^2

Appendix II. Derivation of System Characteristic Equation

Air-Frame Transfer Function

The basic air-frame equations are

$$Z_{\delta E}\delta E = -Z_w w + \dot{w} - uq \quad (8)$$

$$M_{\delta E}\delta E = -M_w w - M_q q + \dot{q} \quad (9)$$

$$a_Z = \dot{w} - uq \quad (10)$$

From equations 8, 9, and 10

$$Z_{\delta E}\Delta E(s) = -Z_w W(s) + sW(s) - uQ(s) - w(0) \quad (11)$$

$$M_{\delta E}\Delta E(s) = -M_w W(s) - M_q Q(s) + sQ(s) - q(0) \quad (12)$$

$$A_Z(s) = sW(s) - uQ(s) - w(0) \quad (13)$$

where

Table III. Hypothetical Flight Conditions

Flight Condition Code*	ω_n	ζ	K_1	K_2	$K = K_1/(\omega_n^2)$†
61.4	0.2975	0.099	−0.0151	0.468	5.3
A	0.3	0.1	−0.01665	0.5	5.56
			$\left(-\frac{1}{60}\right)$		
42.0	0.56	0.203	−0.117	10.56	33.7
B	0.6	0.2	−0.1	10	27.95
22.6	0.987	0.207	−0.793	73.4	75.4
C	1	0.2	−0.8	75	75

*For studies reported herein, the hypothetical flight conditions *A*, *B*, and *C* were assumed. These three flight conditions are approximations of the flight conditions 61.4, 42.0, and 22.6 respectively.
†K is the ratio of a^2 to δE at steady state

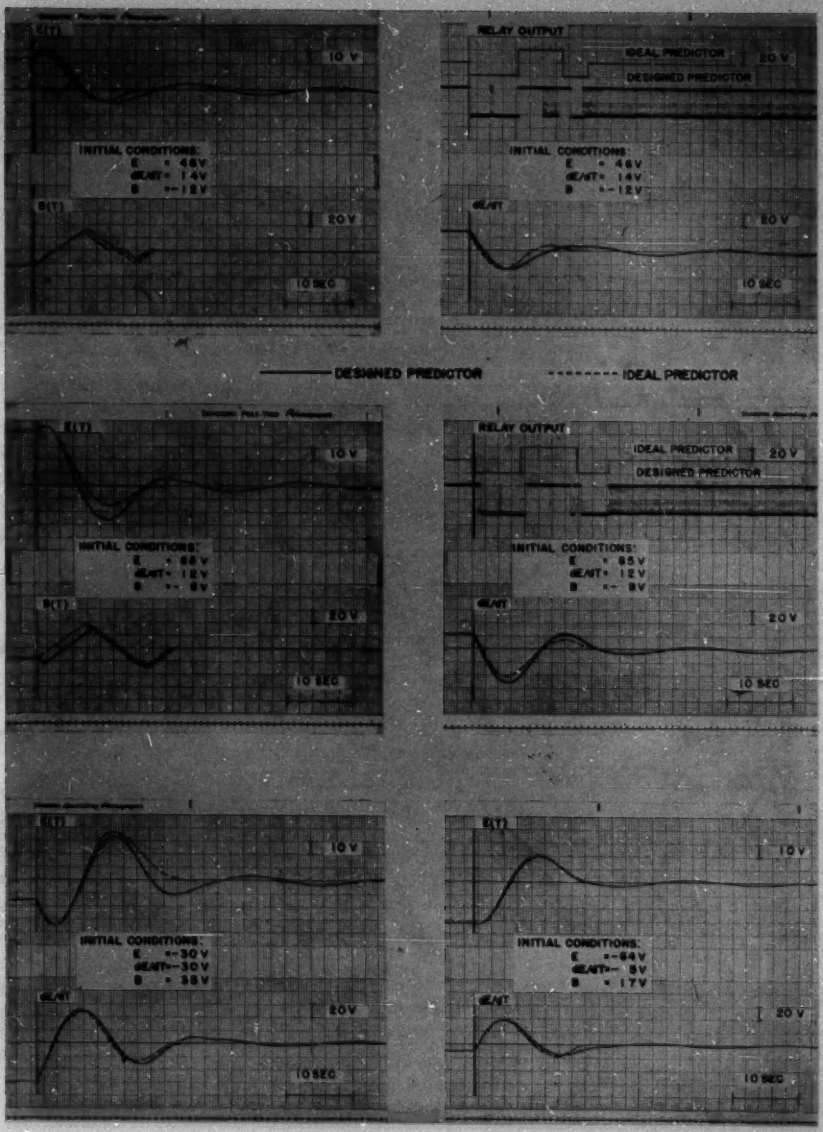


Fig. 11. Comparison of the designed predictor unit and an ideal predictor unit

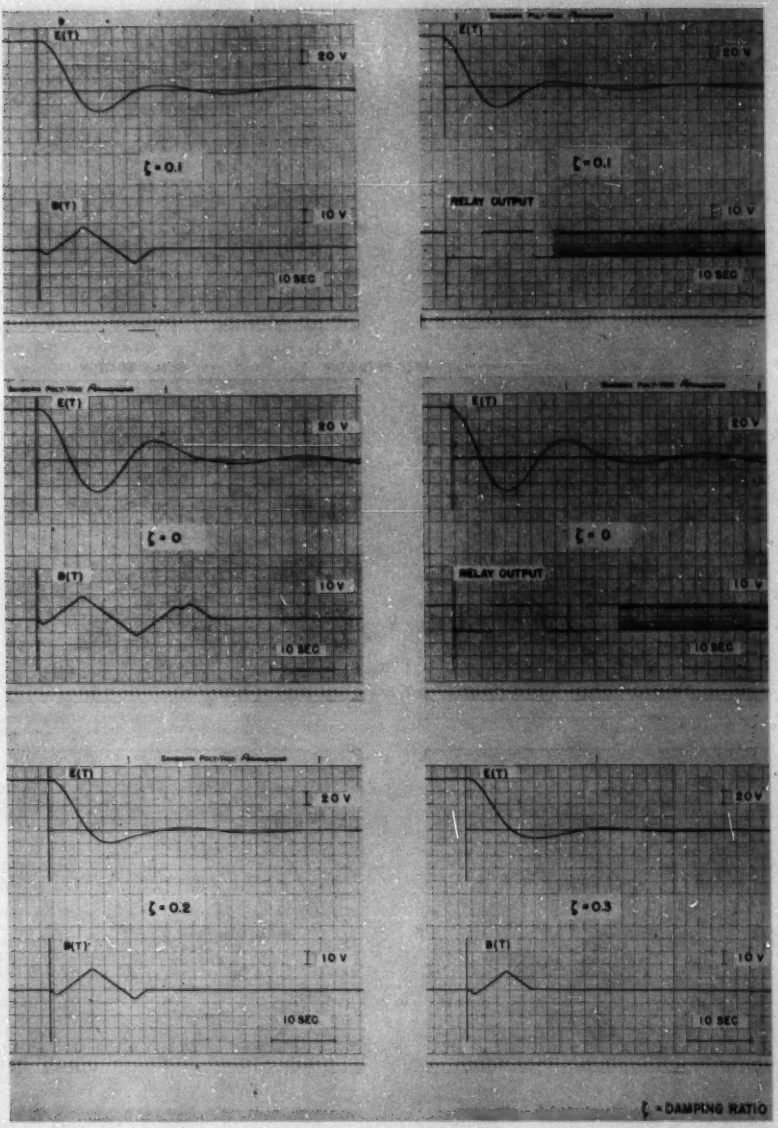


Fig. 12. Effect of change in damping ratio to performance of predictor unit

$A_Z(s) = \mathcal{L}(a_Z)$
$Q(s) = \mathcal{L}(q)$
$\Delta E(s) = \mathcal{L}(\delta E)$
$W(s) = \mathcal{L}(w)$

$$A_Z(s) = \frac{s^2 Z_{\delta E} - s M_q Z_{\delta E} + u(Z_w M_{\delta E} - M_w Z_{\delta E})}{(s - Z_w)(s - M_q) - u M_w} \Delta E + \frac{Z_w[w(0)(s - M_q) + uq(0)]}{(s - Z_w)(s - M_q) - u M_w}$$

For $w(0)$ and $q(0)$ equal to zero

$$\frac{A_Z(s)}{\Delta E(s)} = \frac{s^2 Z_{\delta E} - s M_q Z_{\delta E} + u(Z_w M_{\delta E} - M_w Z_{\delta E})}{s^2 - (M_q + Z_w)s - u M_w + Z_w M_q}$$

$$g_1(s) = \frac{A_Z(s)}{\Delta E(s)} = \frac{s^2 Z_{\delta E} - s M_q Z_{\delta E} + u(Z_w M_{\delta E} - M_w Z_{\delta E})}{s^2 + 2\zeta\omega_n s + \omega_n^2} \quad (14)$$

where

$$\omega_n = \sqrt{Z_w M_q - u M_w} \quad (15)$$

$$\zeta = \frac{-(M_q + Z_w)}{2\sqrt{Z_w M_q - u M_w}} \quad (16)$$

Open-Loop System Transfer Function

If the time lags caused by the actuator and the valve are neglected, the transfer function of the member which precedes the air-frame deflection becomes an integrator with constant gain K_s; see Figs. 2 and 3. The open-loop system transfer function with this approximation is

$$g(s) = \frac{A_Z(s)}{F_c(s)} = K_s \frac{s^2 Z_{\delta E} - s M_q Z_{\delta E} + u(Z_w M_{\delta E} - M_w Z_{\delta E})}{s(s^2 + 2\zeta\omega_n s + \omega_n^2)} \quad (1)$$

System Differential Equation

For the special case where the predictor output $F_c(t)$ is a step of $\pm F_m$ units of amplitude

$$\delta E = \delta E(0) \pm K_s F_m t \quad (17)$$

The differential equation corresponding to equations 14 and 17 form the system differential equation

$$\dddot{a}_Z + 2\zeta\omega_n \ddot{a}_Z + \omega_n^2 \dot{a}_Z = Z_{\delta E}\dot{\delta E} - M_q Z_{\delta E}\delta E + u(Z_w M_{\delta E} - M_w Z_{\delta E})\delta E \quad (18)$$

Substitution of equation 17 into equation 18 yields

$$\dddot{a}_Z + 2\zeta\omega_n \ddot{a}_Z + \omega_n^2 \dot{a}_Z = \pm K_1 K_s F_m + K_2[\delta E(0) \pm K_s F_m t] \quad (19)$$

where

$$K_1 = -M_q Z_{\delta E} \quad (3)$$

$$K_2 = u(Z_w M_{\delta E} - M_w Z_{\delta E}) \quad (4)$$

For $e = a_{Z_c} - a_Z$, equation 19 becomes

$$\dddot{e} + 2\zeta\omega_n \ddot{e} + \omega_n^2 \dot{e} = \omega_n^2 \dot{a}_{Z_c} \mp K_1 K_s F_m - K_2[\delta E(0) \pm K_s F_m t] \quad (2)$$

where

$$\dot{a}_{Z_c}(t) = \dot{a}_{Z_c}(0) \quad (20)$$

a_{Z_c} is assumed to be a step function of time.

The Solution of the Differential Equation of the System Error for a Step Command Signal a_{Z_c}

The solution of equation 2 is assumed to have a form

$$e(t) = a_1 + a_2 t + a_3 e^{-\alpha t}\cos(\beta t + \theta) \quad (21)$$

where

$$\alpha = \zeta\omega_n \quad (22)$$

$$\beta = \sqrt{1 - \zeta^2}\,\omega_n \quad (23)$$

a_1, a_2, a_3, and θ are constants to be evaluated for given initial conditions.

Appendix III. The Adaptation of a Switching Boundary Designed for a Particular Flight Condition to Other Flight Conditions

The system has the characteristic equation of the form

$$\dddot{e} + 2\zeta\omega\ddot{e} + \omega^2\dot{e} = \omega^2\dot{a}_{Z_c} \mp K_1 K_s F_m - K_2[\delta E(0) \pm K_s F_m t] \quad (2)$$

where

ζ, ω, K_1, K_2 = the system parameters and are functions of the operating conditions a_{Z_c}, $\delta E(0)$, K_s, F_m = constants

Let

$$E = m_1 e \quad (24)$$

$$A_{Z_c} = m_2 a_{Z_c}; \quad m_2 = m_1 \quad (25)$$

$$\Delta E = m_3 \delta E \quad (26)$$

$$T = m_4 t \quad (27)$$

Equation 2 becomes

$$\dddot{E} + \frac{2\zeta\omega}{m_4}\ddot{E} + \frac{\omega^2}{(m_4)^2}\dot{E} = \frac{\omega^2}{(m_4)^2}A_{Z_c} \mp \frac{K_1 m_1 K_s F_m}{(m_4)^2} - \frac{K_2 m_1}{m_3 (m_4)^2} \times \left(\Delta E(0) \pm \frac{m_3 K_s F_m T}{m_4}\right) \quad (28)$$

Let the equation for which the switching boundary was designed be

$$\dddot{E} + \frac{2\zeta\omega_0}{m_{40}}\ddot{E} + \left(\frac{\omega_0}{m_{40}}\right)^2 \dot{E} = \frac{\omega_0^2}{(m_{40})^2}A_{Z_c} \mp \frac{K_{10} m_{10} K_s F_m}{(m_{40})^2} - \frac{K_{20} m_{10}}{m_{30}(m_{40})^2} \times \left(\Delta E(0) \pm \frac{m_{30} K_s F_m T}{m_{40}}\right) \quad (29)$$

Equations 28 and 29 are identical if the following relations exist

$$\frac{\omega_0}{m_{40}} = \frac{\omega}{m_4} \quad (30)$$

$$\frac{K_{10} m_{10}}{m_{40}} = \frac{K_1 m_1}{m_4} \quad (31)$$

$$\frac{m_2}{m_4} = \frac{m_{20}}{m_{40}} \quad (32)$$

$$\frac{2\zeta\omega_0}{m_{40}} = \frac{2\zeta\omega}{m_4} \quad (33)$$

$$\frac{K_{20} m_{10}}{m_{30}(m_{40})^2} = \frac{K_2 m_1}{m_3 (m_4)^2} \quad (34)$$

To satisfy all of the foregoing relations, the following must be true

$$\frac{K_1 m_1}{m_4} = \frac{K_{10} m_{10}}{m_{40}} \quad (35)$$

$$\zeta = \zeta_0 \quad (36)$$

In some cases, the terms $(K_1 m_1)/(m_4)$ and $(K_{10} m_{10})/(m_{40})$ are sufficiently small compared to the other terms in equations 28 and 29 and may be neglected. With this assumption, the system characterized by these equations would have the same switching boundary provided the damping ratio ζ is unaltered.

References

1. A Phase-Plane Approach to the Compensation of Saturating Servomechanisms, Arthur M. Hopkin. *AIEE Transactions*, vol. 70, pt. I, 1951, pp. 631–39.

2. Nonlinear Techniques for Improving Servo Performance, Donald McDonald. *Proceedings*, National Electronics Conference, Chicago, Ill., vol. 6, May 1951, pp. 400–21.

3. An Investigation of the Switching Criteria for Higher Order Contactor Servomechanisms, Irving Bogner, Louis F. Kazda. *AIEE Transactions*, vol. 73, pt. II, Jul. 1954, pp. 118–27.

4. Optimization of Nonlinear Control Systems by Means of Nonlinear Feedbacks, R. S. Neiswander, R. H. MacNeal. *Ibid.*, vol. 72, pt. II, Sept. 1953, pp. 262–72.

5. Nonlinear Optimization of Relay Servomechanism, L. M. Silva. Electronic Research Laboratory, University of California, Berkeley, Calif., Apr. 15, 1954.

6. Research in Nonlinear Mechanics as Applied to Servomechanisms, P. E. Kendall, I. Bogner. *Wright Air Development Center Technical Report 53-521*, Wright-Patterson Air Force Base, Cleveland, Ohio, 1953.

A 2-Motor A-C Mine Hoist Control System

A. H. MYLES
MEMBER AIEE

Synopsis: A method for obtaining reduced speed operation of a-c mine hoists is discussed. Control equipment for providing a very slow stable speed for mine shaft and cable inspection and for cable reeving is described. How frequency responsive relays, energized by a frequency generator coupled to one of the motors, control acceleration, slow speed, and overspeed is explained.

THERE are many mine hoists requiring motors ranging from 50 to a few hundred horsepower (hp). It is desirable to drive such hoists with induction motors if satisfactory speed control and safety of operation can be obtained. The hoists may be vertical or slope, and balanced or unbalanced. The load varies over a wide range, often from heavy hoisting to heavy overhauling during one operating cycle.

The basic principle of the control system to be described has long been known, but only recently has it been applied to mine hoists. Originally proposed more than 20 years ago for the hoist motion of large dockside cranes, controllers embodying the basic principle soon found acceptance for shifting the barges beneath such cranes. Now, with frequency-relay control added, controllers using the principle are capable of providing the accurate performance required by a mine hoist.

2-Motor Hoist Drive

The use of two duplicate wound-rotor induction motors provides advantages in continuity of service, reduced WK^2, and ease of control. A balanced, 2-cage vertical hoist powered by two 200-hp 440-volt 3-phase 60-cycle motors is shown in Fig. 1. The motors are geared to a common main shaft. The hoist drums are arranged to be clutched to and unclutched from this shaft. A mine hoist is considered to be a balanced hoist when both drums are clutched to the main shaft and an unbalanced hoist when only one drum is clutched, even though an unbalanced condition can exist when both drums are clutched in. Each motor is equipped with a magnetic service brake, Fig. 2, and each drum has a hydraulically released, gravity-set brake for emergency and manual operation.

Primary and secondary control of the two motors employ double-pole air-break magnetic contactors with associated relays for acceleration and reduced-speed, overspeed, lost-phase, and undervoltage protection; see Fig. 3. The control panel is shown in Fig. 4 before being assembled into its enclosure.

While accelerating and running normally in either direction, the two motors operate in parallel, each supplying half the total hp and torque requirements. Operating speed-torque curves for each motor and maximum unbalanced and maximum normal balanced torque requirements are given in Fig. 5. The torque requirements are shown as vertical broken lines and are in per cent of the full-load torque of one motor with the north cage at the bottom of the mine shaft and loaded. The torque requirements become less as the north cage approaches the top because of the transfer of cable weight. The curves 10 through 17 are for the no. 1 motor when the no. 1 motor is connected for full-speed hoisting of the north cage, and the curves 20 through 27 are for the no. 2 motor when the no. 2 motor is connected for full-speed hoisting of the south cage. Of course the two sets of curves are superimposed when operating with both motors sharing the load. Fig. 5 shows that the load becomes overhauling when being lowered. Consequently, speeds below synchronous speed cannot be accomplished with standard reversing motor control.

To provide proper slowdown and landing speed control for any normal load condition, the torque of one motor is reversed with respect to that of the other motor. Adjustment of the rotor resistance of both the positively and negatively rotating motors gives resultant speed-torque curves which provide the desired slow-speed operation.

Fig. 6 illustrates the derivation of the resultant speed-torque curves by using the 4th master point conditions as an example. When the north cage is loaded and approaching the top, the actual load for an exemplary installation is approximately 12 per cent of the full-load torque of one motor. With the no. 1 motor having a rotor resistance giving speed-torque curve 13, and with the no. 2 motor primary reversed and a rotor resistance giving speed-torque curve 22, the resultant speed-torque curve is *H4*. The torque values of the curve *H4* are the algebraic sums of the corresponding torque values of curves 13 and 22. For example, at approximately 25-per-cent speed, the torque of the no. 1 motor is *T13* and that of the no. 2 motor is *T22*. Subtracting *T22* from *T13* gives *TH4* or 12-per-cent resultant torque in the hoisting direction. Similarly, the resultant curve *L4* is derived from the individual motor speed-torque curves 23 and 12 and indicates the performance when the north cage is being lowered.

By proper selection of the individual motor speed-torque curves for each master control point, a family of desirable resultant speed-torque curves can be obtained, as shown in Fig. 7. This permits the speed under any condition of loading to be reduced in selective steps through a creeping speed to a stalled condition.

If the torque requirements for slowing down the hoist to a stalled condition are greater than those for one motor, then the master points giving the strongest retarding torque can be arranged to provide reverse torque of both motors. With the two motors aiding, adequate deceleration torque can be provided but with some sacrifice in speed regulation with load.

Frequency Responsive Relays

The power supply for the frequency responsive relays is provided by an induction generator of less than 1-kw capacity mechanically coupled to one of the motors through a 1-to-1 gear ratio and having a 3-phase primary and a single-phase secondary. The primary phase rotation is always the same as that of the no. 1 motor. Since the generator is wound with fewer poles per phase then the hoist motors and, therefore, has a higher synchronous speed, the generator rotor voltage and frequency, both of which

Paper 56-84, recommended by the AIEE Industrial Control Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Winter General Meeting, New York, N. Y., January 30-February 3, 1956. Manuscript submitted December 23, 1953; made available for printing November 18, 1955.

A. H. MYLES is with the Electric Controller and Manufacturing Company, Cleveland, Ohio.

The author wishes to acknowledge the assistance obtained from the papers listed in the references.

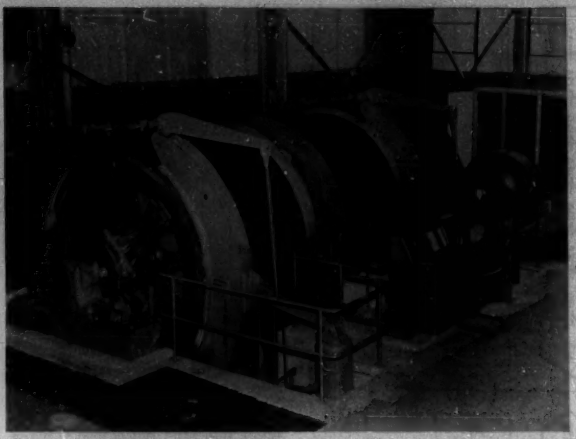


Fig. 1. A 400-hp 2-motor balanced Nordberg hoist showing control desk, centrifugal switches, and control panel

vary inversely with speed, are not zero at the synchronous speed of the hoist motors. Frequency responsive relays of the series resonant and series-parallel resonant type,[1] supplied from the generator, control the acceleration, prevent overspeeding, and provide for an approximately constant creeping speed control. The frequency responsive relay circuits are shown at the lower right in Fig. 3.

Protective Devices and Control Circuits

In all mine hoist installations, particularly those carrying men, many protective devices must be used to assure the safety of personnel and equipment. Combined centrifugal and cam switches geared to each hoisting drum, shown in Fig. 1, provide protection against overspeed and overwind for both directions of travel. Track-type limit switches provide for stopping at the final limit of cage travel. Brake-wear limit switches insure that the drum brakes have adequate stopping ability at all times. Hand-brake lever and clutch position switches provide interlocking protection against releasing a drum brake when its associated drum is unclutched. Relays are provided to assure that an overload, a reverse phase condition, or undervoltage, cause power to be removed and all brakes to set. Fig. 8 is an elementary diagram of the connections for these and other protective devices. It will be noted that backout switches protect the hoist against travel in the wrong direction after operation of the overtravel or overwind limit switches. For example, after either or both the north overwind or the north final limit switches have tripped, the north backout switch must be moved to north backout position before the undervoltage relay *UV* or brake contactor *B* can be reclosed. With this switch in the north backout position, the north up-directional contactor circuit Fig. 9, is opened, allowing operation in the south up-direction only. A circuit breaker is usually provided for additional fault protection and can be operated from the control desk to serve as a disconnect switch.

Normal Speed Operation (800 Feet per Minute)

For hoisting either cage normally, the operation of the controller is as follows:

1. With all knife switches, the circuit breaker, the brake lever switch, and the proper backout switch closed, and the main master switch in the OFF position, the relays *UV* and *B* are energized and complete their own holding circuits. The relay *B* provides power to the brake solenoids for releasing the drum brakes when the respective drums are in their clutched-in position.

2. Moving the main master switch to the first speed point energizes the main primary contactor *M*. If the movement is in the north hoist direction, the directional contactors *1UN* and *2UN* close; if in the south hoist direction, the directional contactors 1*US* and 2*US* close; see Figs. 8 and 9. The motor brakes are released and both motors start with their torques additive. As soon as the final or overwind limit switches reclose, the backout switch is released and returns automatically to its normal position.

3. Moving the main master switch to the second, third, and fourth points successively causes the accelerating contactors to reduce the secondary resistance which increases the

Fig. 2. Two 200-hp motors with electric brakes and control-frequency-generator driving Nordberg mine hoist

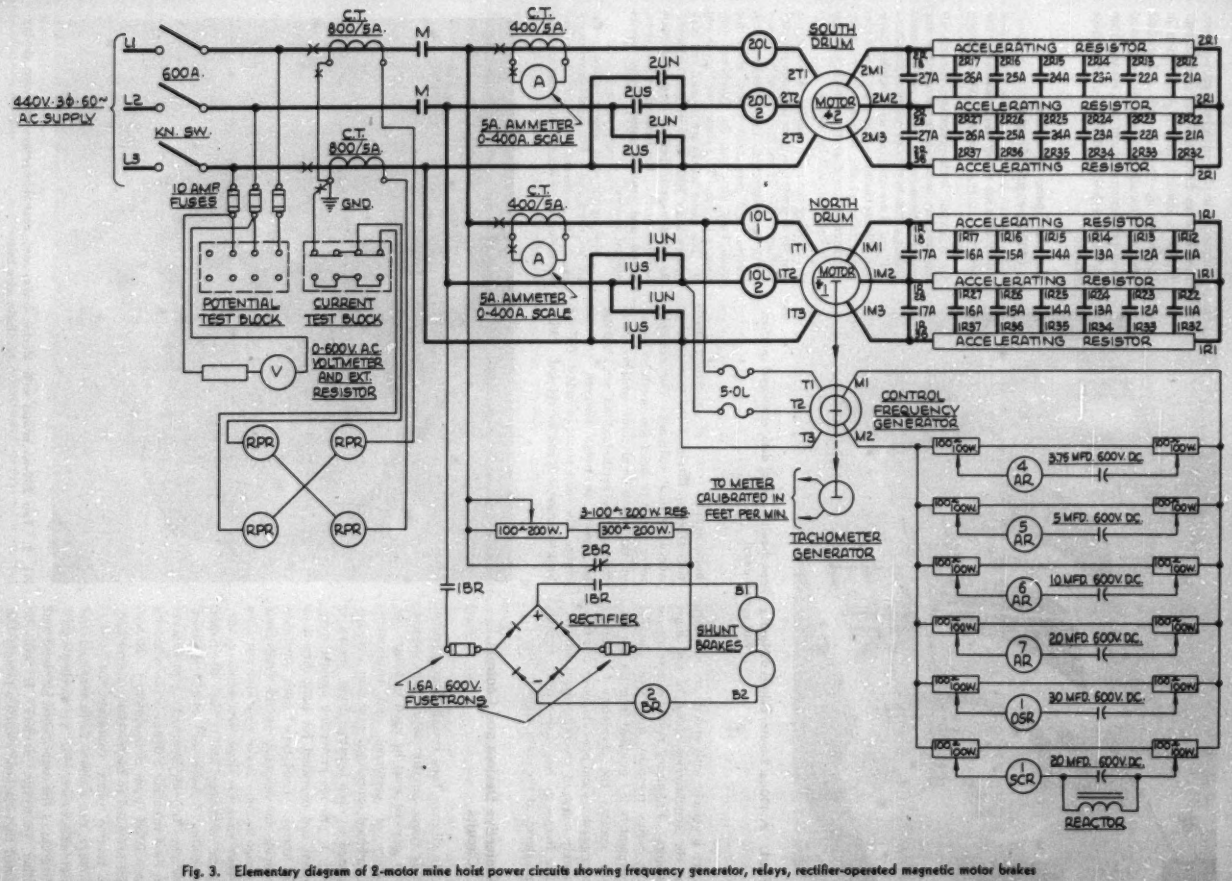


Fig. 3. Elementary diagram of 2-motor mine hoist power circuits showing frequency generator, relays, rectifier-operated magnetic motor brakes

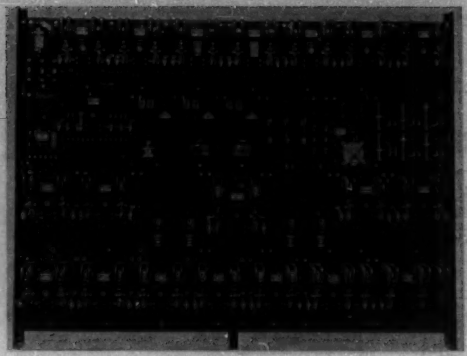


Fig. 4. Control panel for two 200-hp 440-volt mine hoist motors before mounting in enclosure

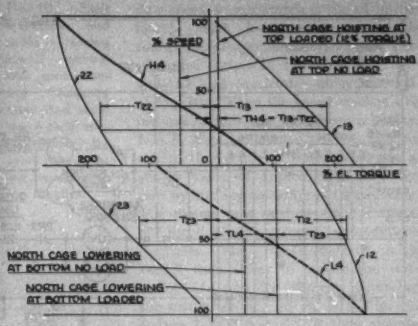

NORTH CAGE HOISTING AT TOP LOADED (12% TORQUE)
NORTH CAGE HOISTING AT TOP NO LOAD
% SPEED
% FL TORQUE
NORTH CAGE LOWERING AT BOTTOM NO LOAD
NORTH CAGE LOWERING AT BOTTOM LOADED

Fig. 6. Method of combining speed-torque curves (fourth master point hoist and lower)

torque of both motors. The hoist accelerates. With the main master switch in the fifth or sixth point, the accelerating relays, which opened their contacts when power was first applied, reclose in succession as the generator rotor frequency and voltage decreases, cutting out the remaining steps of secondary resistance and bringing the motors up to full speed.[1]

4. As the loaded cage nears the top of the shaft, the motor load becomes materially less and the torque must be reduced. The torque is reduced by depressing a push button in the master handle. This energizes the relay *CR* which locks itself in until power is removed. Operation of the relay *CR* causes secondary resistance to be inserted in the rotor circuit of both motors, Figs. 8 and 9, and causes the primary of one motor to be reversed. Operation is now on curve no. 6 of Fig. 7. Moving the master switch from the sixth point back toward the OFF point reduces the torque of the hoisting motor and increases the torque of the reversed motor by successive changes in the secondary resistances. If the north cage is assumed to be hoisting, a stalled-torque condition exists on the third master position; see curve 3 of Fig. 7. After the cages stop, the master switch is returned to the OFF point to remove power and set the brakes.

Operation when lowering a loaded cage is similar except for the final slowdown torque requirements. As the loaded cage moves down, the master switch must be returned to the second or first point lowering to stop the cage; see dotted curves 1 and 2, Fig. 7. The master is then moved to the OFF point to remove power and set the brakes.

Maximum Torque Switch

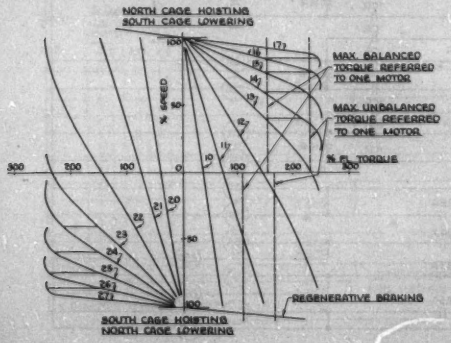
When maximum torque is required to start the cages, the foot switch shown in Fig. 9 is depressed causing the contactors *14A*, and *24A* to reduce the resistance of both motor secondaries. With *14A* and *24A* closed, the rotor resistances and reactances are approximately equal and near maximum stalled torque is developed by both motors; curve 14, Fig. 5.

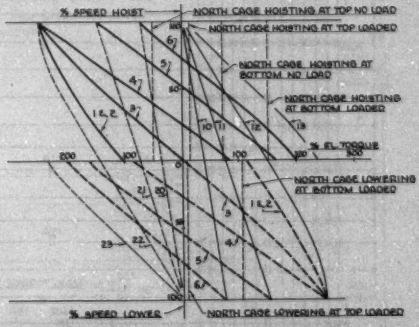
Overspeed Protection

If, owing to an overhauling load when on an intermediate master point, the cage speed should exceed that corresponding to the synchronous speed of the motors, relay *1-OSR* drops out to close contactors *17A* and *27A*. This removes

Fig. 5 (below). Characteristic speed-torque curves for each wound-rotor motor of 2-motor mine hoist

Fig. 7 (right). Deceleration and slow-landing speed performance curves obtained with control of 2-motor hoist




% SPEED HOIST
NORTH CAGE HOISTING AT TOP NO LOAD
NORTH CAGE HOISTING AT TOP LOADED
NORTH CAGE HOISTING AT BOTTOM NO LOAD
NORTH CAGE HOISTING AT BOTTOM LOADED
% FL TORQUE
NORTH CAGE LOWERING AT BOTTOM LOADED
% SPEED LOWER
NORTH CAGE LOWERING AT TOP LOADED

NORTH CAGE LOWERING		NORTH CAGE HOISTING	
MASTER POINT RESULTANT CURVE	MOTOR SPEED – TORQUE CURVES	MASTER POINT RESULTANT CURVE	MOTOR SPEED – TORQUE CURVES
1 & 2	20 & 12 BUCKING	1 & 2	10 & 22 BUCKING
3	22 & 12 "	3	12 & 22 "
4	23 & 12 "	4	13 & 22 "
5	23 & 11 "	5	13 & 21 "
6	23 & 10 "	6	13 & 20 "

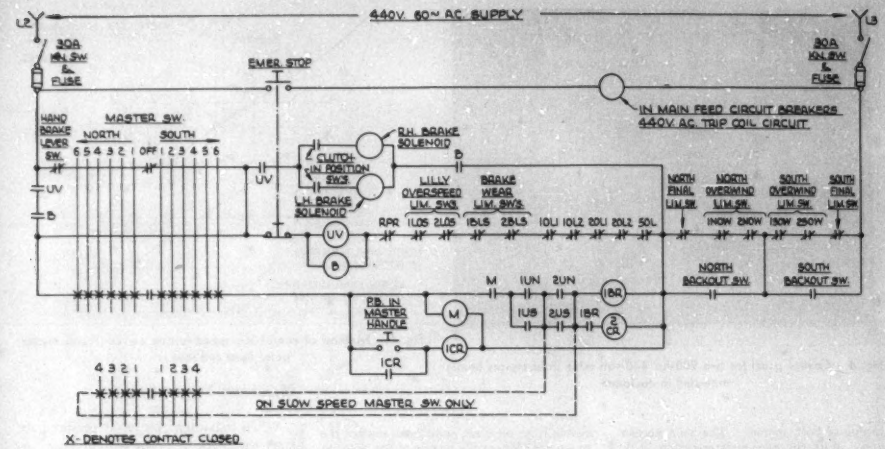
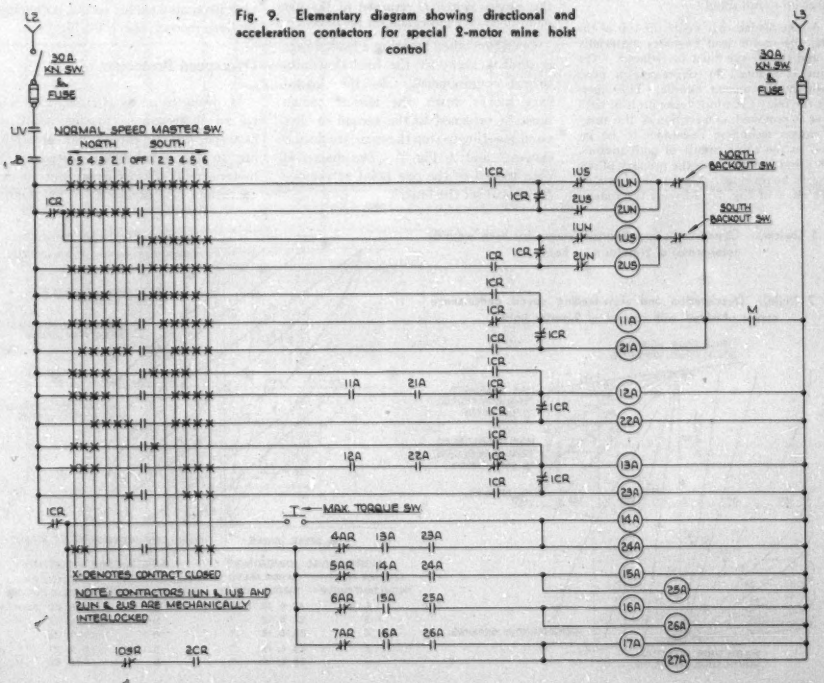


Fig. 8. Elementary diagram of mine hoist safety equipment, interlocking circuits, and push button for slowdown control

Fig. 9. Elementary diagram showing directional and acceleration contactors for special 2-motor mine hoist control



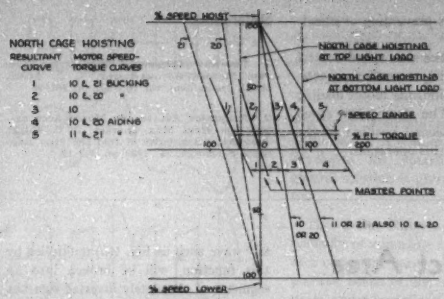

% SPEED HOIST
NORTH CAGE HOISTING
RESULTANT CURVE
MOTOR SPEED-TORQUE CURVES
1 10 & 21 BUCKING
2 10 & 20 "
3 10
4 10 & 20 AIDING
5 11 & 21 "
NORTH CAGE HOISTING AT TOP LIGHT LOAD
NORTH CAGE HOISTING AT BOTTOM LIGHT LOAD
SPEED RANGE
% 21 TORQUE 200
MASTER POINTS
10 OR 20
11 OR 21 ALSO 10 & 20
% SPEED LOWER

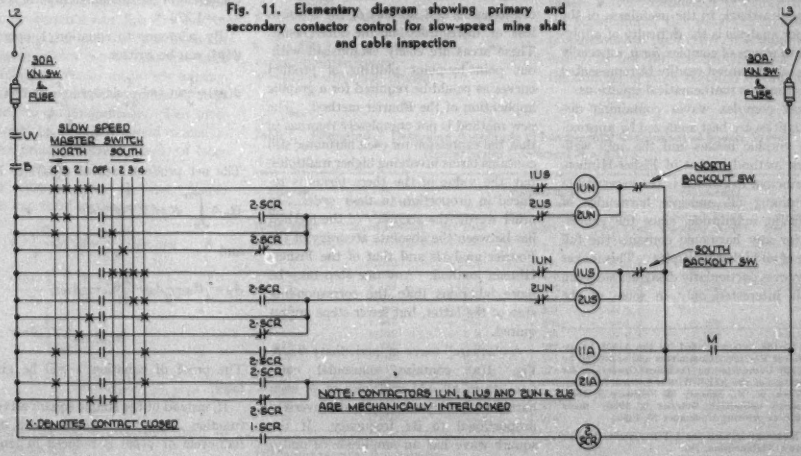
Fig. 10. Shaft inspection and cable - reeving speed control with speed-control frequency relay and separate 4-point master switch; the south cage hoisting operation is identical

all of the secondary resistance and prevents the motors from overspeeding. The *1-OSR* relay contacts remain closed until the master is returned to the OFF point or until the push button is depressed for slowdown operation.

Slow-Speed (100 Feet per Minute) Control for Cable and Shaft Inspection

In all mine hoist installations a relatively constant slow speed should be provided for shaft and cable inspection. The inspection speed usually desired is from 75 to 110 feet per minute or about 12 per cent of synchronous speed for the hoist described. Since an operating cycle at this slow speed is often of considerable duration, and since natural ventilation is reduced because of the slow speed, the motors should not be excessively loaded. By automatically controlling the rotor resistance and the primary connections of the motors, the slow stable speed can be obtained with the 2-motor drive. Figs. 10 and 11 show how the frequency responsive relay *1-SCR*, which has a close differential between pickup and dropout, holds the speed at approximately 100 feet per minute throughout the entire travel distance. A separate 4-speed master switch is used to control the primary connections of the motors. Each speed point provides motor connections which permits the relay *1-SCR* to maintain control for a certain portion or zone of the travel distance. As the cages travel out of a control zone, the relay *1-SCR* loses control until the master is moved to the next point. A transfer switch is used to transfer from the main to the slow-speed master switch. Fig. 11 shows the control circuits involved in the slow-speed control. When an empty cage is at the bottom, the torque required to hoist it is approximately 95 per cent of the full-load torque of one motor. If the slow-speed master switch is placed in the fourth point hoist, each motor is connected for approximately 75 per cent of stalled torque in the same direction and the load accelerates along curve 5 of Fig. 10. When the speed reaches 100 feet per minute, the speed-control relay *1-SCR*, Fig. 3, operates relay *2-SCR* to reduce the torque of both motors to a stalled value of approximately 35 per cent by opening contactors *11A* and *21A*. The motors now operate along curve 4 of Fig. 10. The speed immediately decreases. The speed-control relay *1-SCR* then closes to reclose the contactors *11A* and *21A*, thereby increasing the motor torque; see curve 5, Fig. 10. Relays *1-SCR* and *2-SCR* thus operate repeatedly to hold the speed to near 100 feet per minute. As the cage hoists, the torque required becomes less and eventually the speed-control relays lose control, as shown in Fig. 10. The slow-speed master must then be moved to the third point which results in the no. 2 motor being disconnected from the line, and the speed-control relays *1-SCR* and *2-SCR* again take over to control the torque of the no. 1 motor only by changing its secondary resistance. As the cage continues to hoist and the load becomes still less, the master must be moved to the second point. The speed control relays now connect the no. 2 motor for weak down torque, when required, to hold the speed to 100 feet per minute. As the cage approaches the upper limit, the hoist becomes overhauling in the up direction, and the master must be moved to the first point. On the first point, the

Fig. 11. Elementary diagram showing primary and secondary contactor control for slow-speed mine shaft and cable inspection



speed-control relays control the value of down torque on no. 2 motor to hold the speed at 100 feet per minute. When the upper limit is reached, the master is returned to OFF position to remove power and set the brakes. Under all slow-speed operation, the torque requirements of the motors are 70 per cent of rated torque or less, thus preventing overheating.

Another use of the slow speed control is to facilitate reeving or unreeving cable. During this operation, the motor loading is friction only. The steady slow speed obtained on the second position of the master is ideal for this operation.

References

1. TWO NEW METHODS OF ACCELERATING ELECTRIC MOTORS AUTOMATICALLY, J. D. Leitch. *AIEE Transactions*, vol. 60, 1941, pp. 487–93.

2. FREQUENCY RELAYS AND MOTOR ACCELERATION FOR STEEL MILL CONTROL, D. C. Wright. *Proceedings*, Association of Iron and Steel Engineers, Pittsburgh, Pa., 1939, pp. 302–11.

Harmonic Analysis by Direct Area Measurements

I. K. DORTORT
MEMBER AIEE

ANY PERIODIC function can be expressed as a Fourier series, a summation of sinusoidal terms having frequencies which are integral multiples of the fundamental frequency. The most direct method of determining the coefficient of any term consists basically of determining the net area, over one full cycle, of a curve which is the product of the function and a unit sine-wave of the desired frequency. No matter what the appearance of the product curve, the net area will be exactly the same as that which would be produced by the in-phase component of the desired harmonic multiplied by a unit sine wave of the same frequency. The procedure is repeated with a unit cosine wave and the two components combined in quadrature.

The drawback to the usefulness of the Fourier analysis is the difficulty of applying it to waves of complex form, especially those which cannot readily be represented by continuous mathematical equations.

Most complex waves containing discontinuities are best analyzed by approximate graphic means and the only well-known method is that of Fisher-Hinnen. Its principal drawback is the necessity of determining all multiple harmonics of significant magnitude, since the expression for any harmonic contains the full value of all higher multiples. This makes the process particularly disagreeable when one is interested only in some of the harmonics but the wave under analysis contains an infinite series.

The voltages and currents existing in the primaries of rectifiers are good examples of this type of wave. Since the Fisher-Hinnen method makes use of ordinates of the composite wave, spaced at discrete regular intervals, irregularities of the curve falling between these ordinates will not be represented in the expression for a particular harmonic. For this reason the process must be started with the highest significant harmonic unless one has previous knowledge of the wave, to make sure that no appreciable error exists in the determination of the harmonic of particular interest.

The method described here is based on a direct measure of areas instead of ordinates so that no part of the curve is lost or disregarded in the final result. These areas are easily established without point-by-point plotting of product curves as would be required for a graphic application of the Fourier method. The new method is not completely rigorous in that the expression for each harmonic still contains terms involving higher multiples, but the value of the these terms is reduced in proportion to their order. In other words, the accuracy of the method lies between the absolute accuracy of the Fourier analysis and that of the Fisher-Hinnen method. Any one step may be more laborious than the corresponding step of the latter, but fewer steps are required.

A "square" wave of periodicity $2\pi/p$, Fig. 1(a) contains sinusoidal components in all odd multiples of p, each harmonic having an amplitude inversely proportional to its frequency. If the square wave has an amplitude of unity, any wave such as Fig. 1(b) multiplied by this function, will be broken into $2p$ segments of alternately inverted sign but with exactly the same shape and magnitude as the corresponding segments of the original wave, Fig. 1(c). This unit sqaure wave operator will be called $\psi(p\theta)$.

The net area obtained by integrating the resultant wave from 0 to 2π will be a measure of those harmonics which are common to the original wave and $\psi(p\theta)$. Since each harmonic present in $\psi(p\theta)$ is reduced by its order above p, the area so found will be made up mostly of the harmonic p with progressively reduced amounts of the higher harmonics.

A square wave having a magnitude of unity can be represented by the Fourier series:

$$F(\theta)=A_1\sin\theta+A_3\sin 3\theta+A_5\sin 5\theta\ldots+A_{n'}\sin n'\theta$$

$$=\frac{4}{\pi}\sin\theta+\frac{4}{3\pi}\sin 3\theta+\frac{4}{5\pi}\sin 5\theta\ldots.+\frac{4}{n'\pi}\sin n'\theta \quad (1)$$

n' being any positive odd integer. The coefficient of the fundamental term is then $4/\pi=1.273$.

By reference to equation 1, operator $\psi(p\theta)$ can be written

$$\psi(p\theta)=\frac{4}{\pi}\sin p\theta+\frac{4}{3\pi}\sin 3p\theta+\frac{4}{5\pi}\sin 5p\theta\ldots.+\frac{4}{n'\pi}\sin n'p\theta \quad (2)$$

The net product area of $\psi(p\theta)F(\theta)$ is

$$\mathfrak{A}_p=\int_0^{2\pi}\psi(p\theta)\times F(\theta)\times d\theta \quad (3)$$

and

$$A_p=\frac{1}{4}\mathfrak{A}_p-\frac{1}{3}A_{3p}-\frac{1}{5}A_{5p}-\frac{1}{7}A_{7p}\ldots\ldots-\frac{1}{n'}A_{n'p} \quad (4)$$

The proof of equation 4 will be given later.

If, instead of the simple square wave, a function $M_m(p\theta)$ is used in which each half-cycle of $\psi(p\theta)$ is reduced in length

Paper 56-188, recommended by the AIEE Industrial Power Rectifiers Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Winter General Meeting, New York, N. Y., January 30-February 3, 1956. Manuscript submitted October 3, 1955; made available for printing November 23, 1955.

I. K. DORTORT is with the I-T-E Circuit Breaker Company, Philadelphia, Pa.

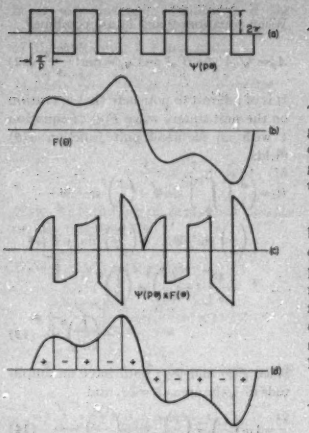


Fig. 1. a—unit square wave operator ($p\theta$), b—wave to be analyzed $F(\theta)$, c—$F(\theta) \times \psi(p\theta)$, d—simplified method of constructing $\int F(\theta) \times \psi(p\theta)$ for measurement

by π/mp, with half of that amount chopped off each end as shown in Fig. 2(a), the mpth harmonic and all its multiples are eliminated and will also be missing from the resultant wave, Fig. 2(b). By this means the number of steps required to obtain a given accuracy can be greatly reduced. Generally, it will be found desirable to cast out the $3p$th harmonic in this manner. However, in an analysis of a current wave produced in a 3-phase system in which no third harmonic currents can flow, m can be made to equal 5 and the $5p$th harmonic and its multiples can be cast out since the triple-n harmonics are already missing. This further reduces the number of steps required to obtain accuracy. This process should be especially useful in analyzing the primary current waves of large power rectifiers.

The chopped-wave operator $M_3(p\theta)$ can be expressed in the series

$$M_3(p\theta) = \frac{2\sqrt{3}}{\pi}\sin p\theta - \frac{2\sqrt{3}}{5\pi}\sin 5p\theta - \frac{2\sqrt{3}}{7\pi}\sin 7p\theta + \frac{2\sqrt{3}}{11\pi}\sin 11p\theta + \frac{2\sqrt{3}}{13\pi}\sin 13p\theta \ldots + \frac{4}{n'\pi}\cos\frac{n'\pi}{6}\sin n'p\theta \quad (5)$$

Following the same procedure used to obtain equation 4,

$$A_p = \frac{1}{2\sqrt{3}}\mathfrak{A}_p + \frac{1}{5}A_{5p} + \frac{1}{7}A_{7p} - \frac{1}{11}A_{11p} - \frac{1}{13}A_{13p} \ldots - \frac{2}{n'\sqrt{3}}A_{n'p}\cos\frac{n'\pi}{6} \quad (6)$$

The proof of equations 5 and 6 will be given later. A similar equation can be developed for operator $M_3(p\theta)$. The general solution is given under "Derivation."

The Fisher-Hinnen method consists basically of summing up those ordinates of a composite wave which occur at the positive peaks of the harmonic components under investigation. Let the harmonic be of the order p and call the sum of the ordinates Y_p. Then, for the square wave, the amplitude of the pth harmonic is given by

$$A_p = \frac{1}{p}Y_p + A_{1p} - A_{3p} + A_{5p} - A_{7p} + A_{11p} \ldots$$

This equation is stated here for comparison with equations 4 and 6.

Nomenclature

$F(\theta)$ = any periodic function which can be represented by a Fourier harmonic series

$\psi(p\theta)$ = square-wave operator of periodicity $2\pi/p$ radians

$\psi'(p\theta)$ = same as $\psi(p\theta)$ but advanced $\pi/2p$ radians

$M_m(p\theta)$ = chopped square-wave operator in which the mpth harmonic and all multiples are missing

$M_m'(p\theta)$ = same as $M_m(p\theta)$ but advanced $\pi/2p$ radians

θ = independent variable of angular displacement

α_n = phase angle displacement of nth harmonic

p = ratio of principal frequency of operator $\psi(p\theta)$ to fundamental frequency of $F(\theta)$

n = any positive integer

n' = any positive odd integer

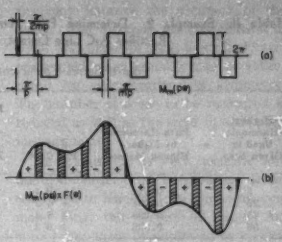


Fig. 2. a—chopped-wave operator $M_m(p\theta)$, b—$\int F(\theta) \times M_m(p\theta)$

m = the lowest order odd harmonic of p missing in operator $M_m(p\theta)$

A_n = Fourier harmonic series coefficients

$\mathfrak{A}_p$ = net area of product curve $\psi(p\theta) \times F(\theta)$

A'_p = approximate peak value of in-phase component of pth harmonic, neglecting higher order terms

$\mathfrak{B}_p$ = net area of product curve $\psi'(p\theta) \times F(\theta)$

B'_p = approximate peak value of quadrature component of pth harmonic, neglecting higher order terms

Y_p = sum of p ordinates when solving for pth harmonic by the Fisher-Hinnen method

Examples

Two examples are presented, with the two operators $\psi(p\theta)$ and $M_3(p\theta)$ used in first, and $M_5(p\theta)$ in the second. The Fisher-Hinnen method was also applied to each problem. The results are compared for accuracy against the true Fourier series coefficients. For simplicity and ease of operation, the first example (Table I) analyzes a symmetrical square wave. Because it is known that the harmonics of this wave are inversely proportional to their order, it is only necessary to test for the fundamental and the number of steps required to obtain a

Table I. Example 1: Determine Amplitude of Fundamental Component of Square Wave
Fourier Series Coefficient $A_1 = 1.273$

Highest Harmonic Used in Given Step*	Fundamental, by Fisher-Hinnen Method	Error, Per Cent	Fundamental, by Square-Wave Operator $\psi(p\theta)$	Error, Per Cent	Fundamental, by Chopped-Square Wave Operator $M_3(p\theta)$	Error, Per Cent
1	1.000	−21.4	1.571	+23.4	1.208	−5.1
3	1.333	+ 4.7	1.396	+ 9.7	†	
5	1.135	−11.0	1.334	+ 4.8	1.256	−1.3
7	1.276	+ 0.2	1.301	+ 2.2	1.281	+0.6
9	1.276	+ 0.2	1.301	+ 2.2	†	
11	1.367	+ 7.4	1.288	+ 1.2	1.271	−0.2
13	1.290	+ 1.3	1.279	+ 0.5	1.264	−0.7
15	1.223	− 3.9	1.286	+ 1.0	†	
17	1.164	− 8.6	1.281	+ 0.6	1.265	−0.4
19	1.216	− 4.5	1.276	+ 0.2	1.271	−0.2
21	1.264	− 0.7	1.280	+ 0.5	†	
23	1.307	+ 2.7	1.277	+ 0.3	1.269	−0.3
25	1.307	+ 2.7	1.277	+ 0.3	1.269	−0.3
27	1.270	− 0.2	1.279	+ 0.5	†	

*No even harmonics present.
†Harmonics cancelled in integration.

Table II. Example 2: Determine Fifth Harmonic of Synthesized 6-Phase Double-Way Rectifier A-C Line Current, 24-Degree Overlap (Fig. 3)

Fourier Series Coefficient $I_5 = \sqrt{A_5^2 + B_5^2} = 0.192$

Highest Harmonic Used in Given Step	Fifth Harmonic, by Fisher-Hinnen Method	Error, Per Cent	Fifth Harmonic by Chopped-Wave Operator $M_5(p\theta)$, Using			
			Planimeter	Error, Per Cent	Estimated Areas	Error, Per Cent
5	0.212	+10.4	0.193	+0.5	0.196	+2.1
25	0.204	+ 6.2			*	
35	0.208	+ 8.3			0.194	+1.0
55	0.208	+ 8.3			0.194	+1.0

* Harmonics cancelled in integration.

given accuracy for the fundamental will then be the same, theoretically, as for obtaining that accuracy for any of the harmonics.

In the second example (Table II) the synthesized wave of the a-c line current of a 6-phase double-way rectifier was analyzed for the fifth harmonic (see Fig. 3). Areas were determined by visual estimate of a plot on 1/10-inch cross-section paper. A half-cycle was made 10 inches long and the amplitude 5 inches high. A single step measurement with a planimeter was also made of the fifth harmonic, disregarding all higher harmonics. The error was only 0.5 per cent. The next harmonic to enter into the result would have been the 35th, and its effect would only be on the order of (1/7)² or 2 per cent. The error in using the planimeter for 35 slivers of areas would probably be greater than 2 per cent, and this method was not pursued further.

Derivation

The general form of the Fourier series is

$$F(\theta) = A_0 + A_1 \sin(\theta + \alpha_1) + A_2 \sin(2\theta + \alpha_2) + A_3 \sin(3\theta + \alpha_3) \ldots + A_n \sin(n\theta + \alpha_n) \quad (7)$$

For the purpose of this paper, A_0 will be considered as eliminated by proper adjustment of the zero line.

It is readily proved that for a unit square wave the coefficients become those given in equation 1 and the displacement angles α_n are zero. Let $\psi(p\theta)$ be a unit function of the same form, but with a periodicity of $2\pi/p$, as shown in Fig. 1(a). This wave is expressed by equation 2.

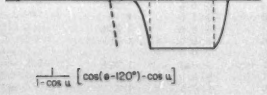


Fig. 3. Synthesized a-c line current of rectifier

Any periodic function $F(\theta)$, multiplied by or operated on by $\psi(p\theta)$, will result in a new series having a constant, or "offset," term for each frequency common to $F(\theta)$ and $\psi(p\theta)$. If the two harmonics are displaced by $\pi/2p$, the offset term is reduced to zero.

The total offset is, of course, the average ordinate of the net area of the resultant function taken over one full cycle. This area is given by

$$\mathfrak{A}_p = \int_0^{2\pi} \psi(p\theta) F\times(\theta) d\theta = \int_0^{2\pi} \sum\left[\frac{4}{n'\pi}\sin n'p\theta\right] \sum [A_n \sin(n\theta + \alpha_n)] d\theta$$

$$= 4 \sum \frac{A_{n'p}}{n'} \cos \alpha_{n'} \quad (8)$$

or

$$A_p \cos \alpha_p = \frac{1}{4}\mathfrak{A}_p - \sum_3^{\infty} \frac{A_{n'p}}{n'} \cos \alpha_{n'p} \approx A_{p'} \quad (9)$$

A'_p indicates an approximation obtained by neglecting the second term on the right-hand side entirely, or carrying the series only to a limited number of terms If $\alpha_{n'p} = 0$ for all values of n', equations 4 and 9 are identical.

Let $\psi'(p\theta)$ be the same function as $\psi(p\theta)$ but advanced $\pi/2p$ radians. Then $\psi'(p\theta) = \psi(p\theta) \sin n'\pi/2$. If the net product area is called $\mathfrak{B}_p$,

$$\beta_p = 4 \sum \frac{A_{n'p}}{n'} \sin n' \frac{\pi}{2} \sin \alpha_{n'p} \quad (10)$$

and

$$A_p \sin \alpha_p = \frac{1}{4}\beta_p - \sum_3^{\infty} \frac{A_{n'p}}{n'} \sin n' \frac{\pi}{2} \sin \alpha_{n'p} \approx B'_p \quad (11)$$

When equations 9 and 11 are combined,

$$A_p = \sqrt{A'^2_p + B'^2_p} \text{ and } \alpha_p = \tan^{-1} \frac{B'_p}{A'_p} \quad (12)$$

It is of interest to note here that operation on the unit square wave $F(\theta)$ of equation 1 with an identical unit function $\psi(\theta)$ yields

$$\mathfrak{A}_1 = \left(\frac{4}{\pi}\right)^2 \int_0^{2\pi} \left[\sin^2\theta + \left(\frac{1}{3}\right)^2 \sin^2 3\theta + \left(\frac{1}{5}\right)^2 \sin^2 5\theta \ldots + \left(\frac{1}{n'}\right)^2 \sin^2 n'\theta\right] d\theta$$

$$= \frac{4^2}{\pi}\left[1 + \left(\frac{1}{3}\right)^2 + \left(\frac{1}{5}\right)^2 \ldots + \left(\frac{1}{n'}\right)^2\right] \quad (13)$$

Since $F(\theta)$ and $\psi(\theta)$ both have an amplitude of 1, the area $\mathfrak{A}_1 = 2\pi$, and

$$\frac{\pi^2}{8} = 1 + \left(\frac{1}{3}\right)^2 + \left(\frac{1}{5}\right)^2 \ldots + \left(\frac{1}{n'}\right)^2 \quad (14)$$

This is a known series and furnishes somewhat unnecessary but interesting proof of the work this far.

It one takes a unit square wave $\psi(\theta)$, but chops off each half-cycle, a piece equal to π/m, as shown in Fig. 4, one eliminates from the Fourier series for the new function the term having a frequency m and all terms having frequencies which are multiples of m.

When chopped square wave of Fig. 4 $M_m(\theta)$ is multiplied by $\sin n'\theta$ to determine the Fourier series coefficients, the fundamental has an amplitude

$$A_1 = \frac{2}{\pi}\int_{\pi/2m}^{\pi - \pi/2m} \sin^2\theta d\theta = \frac{4}{\pi}\cos\frac{\pi}{2m} \quad (15)$$

From considerations of symmetry, no even harmonics can exist; therefore:

$$A_{n'} = \frac{2}{\pi}\int_{\pi/2m}^{\pi - \pi/2m} \sin n'\theta d\theta = \frac{4}{n'\pi}\cos\frac{n'\pi}{2m} \quad (16)$$

and

$$M_m(\theta) = \frac{4}{\pi}\cos\frac{\pi}{2m}\sin\theta + \frac{4}{3\pi}\cos\frac{3\pi}{2m}\sin 3\theta + \frac{4}{5\pi}\cos\frac{5\pi}{2m}\sin 5\theta \ldots + \frac{4}{n'\pi}\cos\frac{n'\pi}{2m}\sin n'\theta \quad (17)$$

Obviously, if n' is a multiple of m, that term disappears. Any term in which $4n+1 < n'/m < 4n+3$ will be negative (n

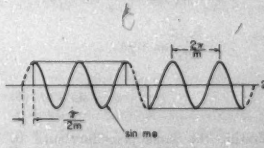


Fig. 4. $M_m(p\theta) \times \sin m\theta$

being any integer). Similarly, if the periodicity of M is $2\pi/p$, the reduction of the half-cycle is π/mp, and

$$M_m(p\theta)=\frac{4}{\pi}\cos\frac{\pi}{2m}\sin p\theta+\frac{4}{3\pi}\cos\frac{3\pi}{2m}\sin 3p\theta+\frac{4}{5\pi}\cos\frac{5\pi}{2m}\sin 5p\theta\ldots.+\frac{4}{n'\pi}\cos\frac{n'\pi}{2m}\sin n'p\theta \quad (18)$$

Now, with the use of $M_m(p\theta)$ to operate on any function $F(\theta)$

$$G_p=\int_0^{2\pi}M_m(p\theta)\times F(\theta)d\theta=\int_0^{2\pi}\sum\left[\frac{4}{n'\pi}\cos\frac{n'\pi}{2m}\sin n'p\theta\right]\times\sum[A_n\sin(n\theta+\alpha_n)]d\theta$$

$$=4\sum\frac{A_{n'p}}{n'}\cos\frac{n'\pi}{2m}\cos\alpha_{n'p} \quad (19)$$

or

$$A_p\cos\frac{\pi}{2m}\cos\alpha_p=\frac{1}{4}G_p-\sum_3^{\infty}\frac{A_{n'p}}{n'}\cos\frac{n'\pi}{2m}\cos\alpha_{n'p}\approx A'_p\cos\frac{\pi}{2m} \quad (20)$$

If $\alpha_{n'p}=0$ and $m=3$, equation 20 reduces to equation 6. Similarly, if $M'_m(p\theta)$ is $M_m(p\theta)$ advanced $\pi/2p$ radians $M'_m(p\theta)=M_m(p\theta)\sin n'\pi/2$ and

$$A_p\cos\frac{\pi}{2m}\sin\alpha_p=\frac{1}{4}\beta_p-\sum_3^{\infty}\frac{A_{n'p}}{n'}\cos\frac{n'\pi}{2m}\sin n'\frac{\pi}{2}\sin\alpha_{n'p}\approx B_p\cos\frac{\pi}{2m} \quad (21)$$

Operator $M(p\theta)$ is useful only when $F(\theta)$ contains no harmonic lower than the m th.

Conclusion

Mention was made at the beginning of the derivation that $F(\theta)$ would be assumed to be adjusted so as to reduce to zero any displacement component A_0. This is desirable for greater accuracy but not necessary. When $F(\theta)$ contains even harmonics the labor involved in reducing A_0 to zero may not be justified.

Even though the operators discussed in this paper contain no even harmonic there is no limitation on the wave form to be analyzed since p and m can be odd or even. Furthermore, it is possible, and perhaps desirable for special cases, to develop an operator containing even harmonics of p. It might take the form of a rectangular wave having a positive lobe of amplitude 2 and only half as long as the negative lobe. No work has been done along this line because its usefulness is not now apparent.

Any tricks devised to reduce the labor involved in this method are justified any may improve the accuracy of the result. For instance, the cancellation of equal and opposite areas is strongly recommended when equality can be established by considerations of symmetry. The number of areas to be measured is reduced as well as the error of each such measurement.

The planimeter is very accurate and useful for lower order harmonics. It becomes quite inaccurate when measuring many long, thin slivers of areas. Accuracy can again be improved by cancelling the base areas of such slivers and measuring only the upper parts.

A large plot on cross-section paper provides a good means for visually determining areas with surprisingly small net errors. When the subdivisions become very small, it is generally permissible to determine the area of each sliver by estimating its average amplitude and multiplying by the calculated $\Delta\theta$. When this is done by use of the square wave operator $\psi(p\theta)$, then for large values of p this method and the Fisher-Hinnen Method become almost identical.

It is not necessary to use the same method for each step of the determination of each harmonic. Any or all methods may be used in solving a problem; the proper choice is the one which offers the greatest ease of operation and accuracy at that step of the computation.

Two Types of Zero-Velocity-Error Servomechanisms

EMANUEL LEVINSON
ASSOCIATE MEMBER AIEE

Synopsis: This paper describes two types of zero-velocity-error servomechanisms which exhibit high dynamic accuracy with a minimum bandwidth. The first type is a special integral control system. The second type contains a simple lag circuit in the feedback path. These systems are analyzed and discussed from both theoretical and practical design considerations. A method to aid in the determination of required bandwidth is introduced. A classification of possible zero-velocity-error systems and a method of deriving zero-forced-error systems are also presented.

IN MANY applications of feedback control systems, extremely high dynamic accuracy is required. At the same time, it is desirable to have as low a bandwidth as possible for the following major reasons:

1. Attenuation of input noise.
2. Avoidance of the so-called region of secondary delays and its destabilizing influence.

Paper 56-195, recommended by the AIEE Feedback Control Systems Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Winter General Meeting, New York, N. Y., January 30–February 3, 1956. Manuscript submitted October 20, 1955; made available for printing December 27, 1955.

EMANUEL LEVINSON is with the Sperry Gyroscope Company, Great Neck, N. Y.

At the same time, of course, the system must be a well stabilized one and should present simplicity of its equipment as well.

Two types of feedback control systems which satisfy these requirements in an optimum fashion are described and analyzed in this paper. Both types are zero-velocity-error systems and both types have identical transfer functions with respect to input dynamics. The systems are characterized by a single parameter related to the bandwidth. The first type is a particular case of integral control system. The second type called the feedback lag system makes use of a simple lag with a critical time constant in the feedback path.

Integral control systems have been discussed extensively in the literature and the feedback lag system was described in an article by King.[1] This paper expands significantly on this previous work, by introducing new methods, principles, and concepts.

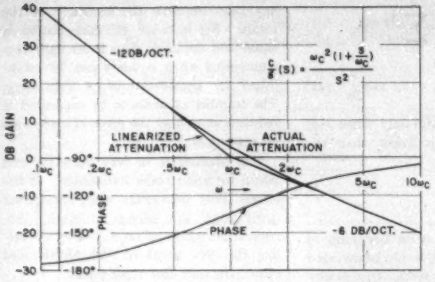


Fig. 1. Db-log ω plot for the -12, -6 crossover system

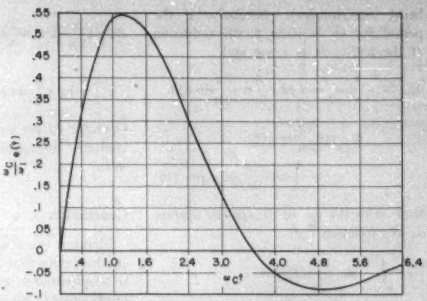


Fig. 3. Normalized error response to step of velocity for the -12, -6 crossover system

The -12, -6 Crossover System

Dynamic Characteristics

Both of the systems to be described in this paper have identical dynamic characteristics with regard to input response. The output-error transfer function is

$$\frac{C}{E}(s) = \frac{\omega_c^2\left(1+\frac{s}{\omega_c}\right)}{s^2} \qquad (1)$$

The linearized decibel (db)-log ω plot of this function in Fig. 1 shows it to have an initial slope of -12 db per octave to the crossover point ω_c and a -6-db-per-octave slope thereafter. The actual attenuation plot is $+3$ db at ω_c. This is a type-2 (zero-velocity-error) system with an acceleration constant $K_A = \omega_c^2$. This system will hereinafter be referred to as the -12, -6 crossover system.

The closed-loop transfer functions are

$$\frac{C}{R}(s) = \frac{1+\frac{s}{\omega_c}}{1+\frac{s}{\omega_c}+\frac{s^2}{\omega_c^2}} \qquad (2)$$

and

$$\frac{E}{R}(s) = \frac{\frac{s^2}{\omega_c^2}}{1+\frac{s}{\omega_c}+\frac{s^2}{\omega_c^2}} \qquad (3)$$

The normalized error response to a unit step displacement input is plotted in Fig. 2

$$e(t) = \frac{2}{\sqrt{3}}\epsilon^{-\omega_c t/2}\sin\left(\frac{\sqrt{3}}{2}\omega_c t + 120°\right) \qquad (4)$$

This response is characterized by a 30-per-cent peak overshoot, a rise time to first crossover equal to $1.2/\omega_c$, and a settling time (to remain within about 95 per cent of input) equal to $4.4/\omega_c$.

The normalized error response to a step velocity input of magnitude ω_t

$$\frac{\omega_c}{\omega_t}e(t) = \frac{2}{\sqrt{3}}\epsilon^{-\omega_c t/2}\sin\left(\frac{\sqrt{3}}{2}\omega_c t\right) \qquad (5)$$

is plotted in Fig. 3. This response is characterized by a zero steady-state error, a peak error equal to 0.55 ω_t/ω_c radians, and a settling time (to within 95 per cent of input position) equal to $6/\omega_c$.

The closed-loop gain and phase frequency responses are plotted in Fig. 4. This response is characterized by a peak gain equal to 1.47 and a bandwidth (unity gain crossover) equal to $\sqrt{2}\omega_c$. The system is thus observed to be a fairly well-stabilized one.

Minimum Peak Velocity Error

It will now be demonstrated that, in a certain sense, the system chosen is an optimum one. The system is the solution to the following problem:

Given a certain bandwidth (defined as the output over error crossover point ω_c) what unconditionally stable system with a phase margin of at least 45 degrees yields a minimum peak error to a step velocity input?

This is shown to be true in Figs. 5 and 6 where the normalized step velocity responses of various -12, -6 and -6, -12, -6 systems are compared. γ and α are the ratios of breakpoint fre-

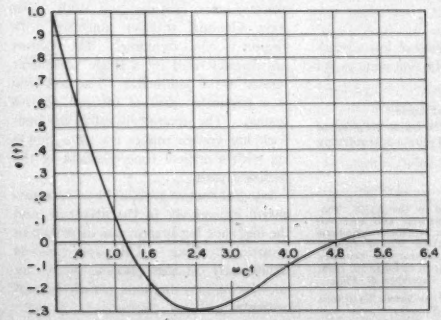


Fig. 2 (left). Normalized error response to step of displacement for the -12, -6 crossover system

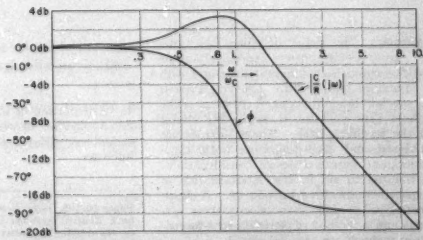


Fig. 4 (above). Frequency response of -12, -6 crossover system

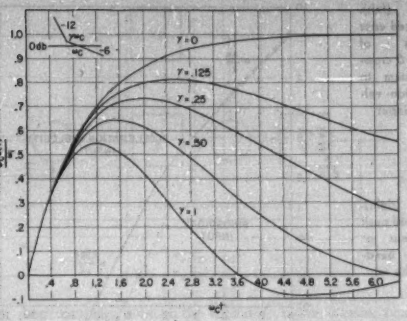


Fig. 5. Velocity error responses for various −12, −6 systems

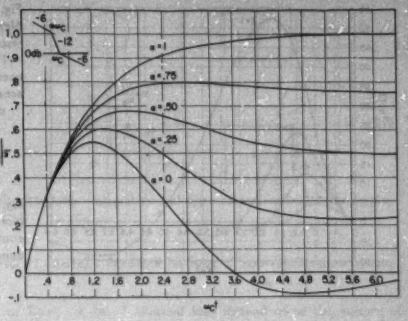


Fig. 6. Velocity error responses for various −6, −12, −6 systems

quencies to crossover frequency, as defined in Figs. 5 and 6. For the conditions stated in this problem, the chosen system also yields the minimum steady-state error to a step acceleration input. Although only −12-db-per-octave maximum-slope systems were compared, it is felt that the optimal condition is still valid when compared with systems such as −6 −18 −6, etc., since the higher slopes can only last over a small frequency range to satisfy the aforementioned constraints.

Bandwidth Determination

The determination of the single system parameter ω_C is generally a best compromise between accuracy, stability (with regard to a forward path lag and inner loop considerations), input noise rejection, and minimum equipment design. As an aid in this determination, the following analysis of the effect of ω_C on dynamic accuracy has been developed.

In many servo applications, a maximum input acceleration α_{max} and a maximum input velocity $\omega_{i\,max}$ are specified. Then, a worst possible input function to be followed is, as shown in Fig. 7, a step of maximum acceleration which lasts until the maximum velocity is reached at which time ($t_0 = 1/\omega_0 = \omega_{i\max}/\alpha_{max}$) the input remains moving at maximum velocity. The error response of the −12 −6 system to this input function has been derived and normalized to

$$\frac{\omega_0}{\omega_{i\,max}} e(t) = \frac{1}{k^2} + \frac{2}{k^2\sqrt{3}} e^{\frac{-k\omega_0 t}{2}} \sin\left(\frac{\sqrt{3}}{2} k\omega_0 t - 120°\right) \quad (6)$$

for $\omega_0 t \leq 1$, and equals

$$\frac{2}{k^2\sqrt{3}} e^{-k\omega_0 t/2}\left[\sin\left(\frac{\sqrt{3}}{2} k\omega_0 t - 120°\right) - k^{1/2}\sin\left(\frac{\sqrt{3}}{2} k\omega_0 t - \frac{\sqrt{3}}{2} k - 120°\right)\right]$$

for $\omega_0 t \geq 1$, where $k = \omega_C/\omega_0$.

This response is plotted in Fig. 8 for various values of k. In Fig. 9, the peak error is plotted versus k on log-log paper. These plots can be used for any values of α and ω_i; the maximum values yield the largest error. For small values of k, the response approaches that for a step of velocity. For large values of k, the response approaches that for a step of acceleration as far as the peak error is concerned. Thus, from this latter plot, given a maximum acceleration and velocity, the value of ω_C required to meet an allowable peak error can be obtained. The following example will illustrate the use of this plot.

Given: $\omega_{i\max} = 1$ radian per second; $\alpha_{max} = 2$ radians per second^2; the allowable peak error is 1 milliradian. Then

$$\omega_0 = \frac{\alpha_{max}}{\omega_{i\,max}} = 2 \text{ radians per second;}$$

$$\frac{\omega_0}{\omega_{i\,max}} e_{peak} = 0.002 \text{ radian}$$

Since this is the region where

$$\frac{\omega_0}{\omega_{i\,max}} e_{peak} = \frac{1.164}{k^2}$$

then

$$\left(\frac{\omega_C}{\omega_0}\right)^2 = k^2 = \frac{1.164}{0.002} = 582$$

and the required value of $\omega_C = 48.4$ radians per second.

It will now be demonstrated that the value of ω_C obtained in the foregoing manner will insure a bandwidth wide enough to pass any steady-state sinusoidal input the magnitude of which is greater than the allowable peak error. Because of the maximum velocity and acceleration constraints, any higher frequency must have a magnitude less than the allowable peak error and, therefore, need not be transmitted.

For a sinusoidal input

$r(t) = \theta_m \sin \omega t$
$\omega_i(t) = \omega\theta_m \cos \omega t$
$\alpha(t) = -\omega^2\theta_m \sin \omega t$

The largest possible ω for $\theta_m \geq \varepsilon$ peak is the value that satisfies both the constraints

$$\omega_m \leq \frac{\omega_{i\,max}}{\varepsilon_{peak}}; \quad \omega_m^2 \leq \frac{\alpha_{max}}{\varepsilon_{peak}} \quad (7)$$

These can be expressed as

$$\frac{\omega_m}{\omega_0} \leq \frac{\omega_{i\max}}{\omega_0 \varepsilon_{peak}}; \quad \frac{\omega_m^2}{\omega_0^2} \leq \frac{\alpha_{max}}{\omega_0^2 \varepsilon_{peak}} = \frac{\omega_{i\,max}}{\omega_0 \varepsilon_{peak}} \quad (8)$$

Therefore, for values of ω_m/ω_0 less than 1, the first constraint is more stringent.

For values of ω_m/ω_0 greater than 1, the second constraint is more stringent. From the frequency response in Fig. 4 it can be observed that any ω_m less than about $1.8\omega_0$ (3-db down point) will essentially be transmitted. The required value of ω_C is then given by

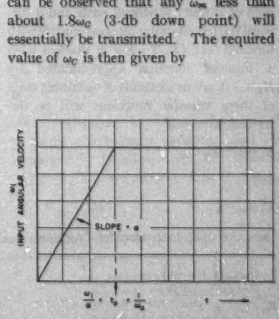


Fig. 7. Acceleration-velocity input function

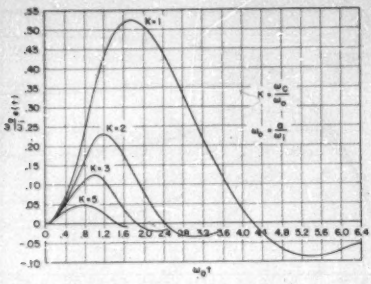
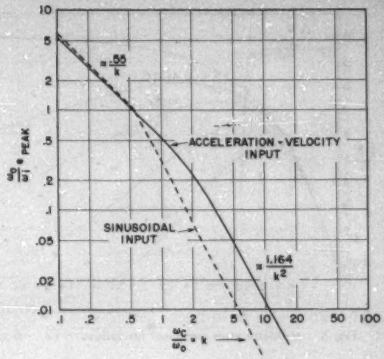


Fig. 8 (left). Normalized error responses of −12, −6 cross-over system to acceleration-velocity input

Fig. 9 (right). Normalized peak error versus required bandwidth



$$k = \frac{\omega_C}{\omega_0} = \frac{\omega_{\epsilon\,\max}}{1.8\omega_{\epsilon\,\text{peak}}}$$

if $k \leq 1/1.8 = 0.555$

$$k^2 = \left(\frac{\omega_C}{\omega_0}\right)^2 = \frac{\omega_{\epsilon\,\max}}{3.24\omega_{\epsilon\,\text{peak}}}$$

if $k \geq 0.555$. These relationships

$$\frac{\omega_{\epsilon\,\text{peak}}}{\omega_{\epsilon\,\max}} = \frac{0.555}{k} \qquad (9)$$

for $k \leq 0.555$, or

$$\frac{0.308}{k^2}$$

for $k \geq 0.555$, are plotted as dashed lines in Fig. 9. It is observed that, except for low values of k where the ω_C values are only slightly different, the ω_C required for the acceleration-velocity input is greater than that required for a sinusoidal input.

Integral Control System

Description of System

The first type of zero-velocity-error servomechanism to be discussed is shown in the block diagram of Fig. 10. It consists of a unity feedback path and a forward path containing integral control with forward feed $(K_1/s+K_2)$ in tandem with a transfer function approximated by K_4/s. Various methods of obtaining each of these transfer functions will be described later. If

$$K_1 = \omega_C^2/K_3$$

and

$$K_2 = \omega_C/K_3$$

then the open-loop transfer function

$$\frac{C}{E}(s) = \frac{\omega_C^2\left(1+\frac{s}{\omega_C}\right)}{s^2}$$

which is dynamically identical with the general −12, −6 single parameter system discussed in the foregoing section. Depending on the physical system employed, one of the K values can be chosen to yield a minimum equipment design. For example, K_3 could be made = 1 (for which case $K_1 = K_2 = \omega_C$), or less than 1 so as not to require any amplification in the forward feed path.

It is well known that a system of this type with an integration between the output and the motor torque yields a zero-torque error to a steady load torque. For this particular system, the error over load torque transfer function, neglecting any forward path lag, can be shown to be

$$\frac{E(s)}{T_L(s)} = \frac{s}{K_1 K_A k_m N\left(1+\frac{s}{\omega_C}+\frac{s^2}{\omega_C^2}\right)} \qquad (10)$$

where

K_A = the gain from the second summing point to motor
k_m = the torque per volt sensitivity of the motor
N = the gear ratio from motor to load

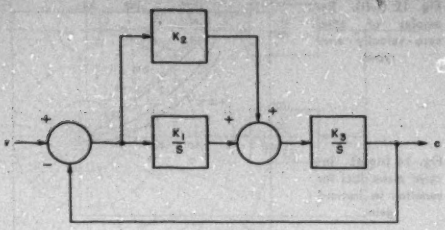
The steady-state error to a constant load torque is observed to be zero. However, some of the electromechanical methods used to obtain the integral control have some static torque of their own which will lead to a static error that is not zero. This will be discussed later. Also, it is still desirable to have a high value of $K_A k_m$ to minimize any dynamic deadzone effects caused by servomotor and load stiction.

Methods of Obtaining K_4/s

If the ω_C value required is small enough and the motor and load time constant is small enough, then the K_4/s can be obtained from the motor integration only. For this case τ_m should be $\leq 1/(8\omega_C)$. For most present 2-phase induction motors for instrument servos, this simple technique can be employed only if $\omega_C \leq 5$ radians per second. To obtain the advantages of equipment simplicity, a compromise with accuracy (lowering the required value of ω_C) might be justifiable.

By adding a direct drag device such as a viscous or eddy-current damper to the motor shaft, the motor time constant can be made much smaller, and much higher ω_C systems can be satisfied. Simple drag damping is analytically equivalent to straight tachometric feedback, which is discussed in the following. Drag damping has the advantages of not requiring a tachometric signal subtraction as well as not having to worry about tachometric null voltage and an inner loop stability. However, drag damping directly subtracts from available power at the load and may force the use of a larger motor. Also, because of the open-loop nature of drag damping, amplifier gain, motor gain, and viscous coefficient variations show up directly in K_4. In viscous fluid dampers, some form of temperature control may be required.

Use of direct tachometric feedback will also decrease the effective motor time constant and K_4 can be made more insensitive to amplifier and motor gain variations. However, for high ω_C, high tachometric loop gain is required to make the effective motor time constant small enough. In the presence of secondary delays (such as magamp lag, d-c motor field lag, or a-c motor-tuned control phase) this high tachometric loop gain may cause instability of the tachometric feedback loop with deleterious effects to the servo. Other secondary delays outside the tachometric loop, such as mechanical resonance, will also limit the value of ω_C possible. In the case of sim-



$K_1 K_3 = \omega_C^2$
$K_2 K_3 = \omega_C$

Fig. 10 (left). Integral control type of zero-velocity-error system

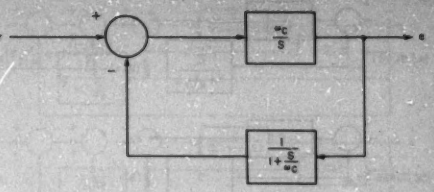


Fig. 11 (above). Feedback lag type of zero-velocity-error system

ple drag damping, none of the secondary delays are reduced and they will provide an upper bound on ω_C. One of the major motivations for advancing these systems is, of course, the avoidance of the region of these secondary delays by obtaining high dynamic accuracy with a minimum bandwidth ω_C.

Methods of Obtaining K_2+K_1/s

Three general methods for obtaining proportional plus integral control will be discussed. These may be classified as electronic, electromechanical, and thermal. A regenerative feedback amplifier with a simple lag in the feedback path[2] can be made to have the desired transfer function by a critical gain adjustment. This adjustment can be held fairly close by use of an internal high-gain feedback amplifier. Possible lag circuits for a-c systems are described in the following section. Various passive and active circuits can be employed to yield an approximation to the desired transfer function. These under compensated integral or lagging doublet circuits have been extensively treated in the literature. The approximation leads to a finite but high K_v system.

One electromechanical method consists of a special motor moving in a highly viscous fluid with a linear synchro pickoff. By thus lowering the motor time constant, an approximation to K_1/s is obtained. This same effect can be obtained by use of a regular motor with tachometric feedback with a miniature gear train and positional pickoff. The comparisons between simple drag and tachometric damping in the discussion for K_1/s are equally valid here.

A second electromechanical method proposed by Cosgriff[3] takes an opposite approach. This method consists of a special high-inertia motor with viscous friction minimized and compensated for. This greatly increases the motor time constant and yields an approximate motor transfer function of K/s^3. A tachometer mounted on the motor shaft is the output device such that the over-all transfer function is K_1/s. An air bearing is used to minimize any stiction and coulomb friction. It is interesting to note that, if these techniques were applied to the main servo motor, then a -12, -6 system could be obtained just by use of a lead network in the forward path.

Each of these electromechanical methods exhibit coulomb friction effects which lead to a steady-state error. The error over torque transfer function can be shown to be

$$\frac{\mathbf{E}(s)}{T_{1m}(s)} = \frac{1}{K_{1A}k_{1m}\left(1+\frac{s}{\omega_C}+\frac{s^2}{\omega_C^2}\right)} \qquad (11)$$

The steady-state error caused by a static torque T_{1m} at the motor shaft of the mechanical filter is observed to be

$$\mathbf{E}_{ss} = \frac{T_{1m}}{K_{1A}k_{1m}} \qquad (12)$$

where

K_{1A} = the gain from the error-measuring point to the control phase of the mechanical filter motor

k_{1m} = the torque per volt sensitivity of that motor

A typical thermal integrator is actually a thermal lag circuit with a very large time constant so as to approximate a pure integration. It consists of a differentially controlled heater bridge with an isolated temperature-sensitive bridge output circuit. The differential heating causes the differential temperature change to be proportional to the input voltage (with a thermal lag) instead of to the square of the input and, together with the output bridge, makes the unit polarity sensitive and theoretically ambient insensitive. As in all bridge circuits, unbalance effects occur which can lead to system error.

Feedback Lag System

Description of System

The second type of zero-velocity-error servomechanism to be discussed is shown in the block diagram of Fig. 11. It consists of a forward path approximated by $G(s)=\omega_C/s$ and a lag in the feedback path having a break point at precisely the same value as the gain of the forward path $H(s)=1/(1+s/\omega_C)$. The forward path transfer function is obtained by use of straight tachometric feedback, a mechanical damper, or the motor transfer function itself. As will be shown, however, the need for a fairly stable gain value practically demands the use of straight tachometric feedback. Various methods of obtaining the feedback lag will be described later.

Substitution in the general expression

$$\frac{C}{\mathbf{E}}(s) = \frac{G(s)}{1+G(s)H(s)-G(s)} \qquad (13)$$

yields

$$\frac{C}{\mathbf{E}}(s) = \frac{\omega_C^2\left(1+\frac{s}{\omega_C}\right)}{s^2}$$

which is the open-loop transfer function of a zero-velocity-error system and is dynamically identical with the general $-12-6$ crossover system. $\mathbf{E}$ is, of course,

Table I. Classification of Zero-Velocity-Error Systems

1.	$n=0$; $m=0$; $1+K_1K_2-K_1=0$	$b_1a_0-a_1a_0+(K_1-1)(a_{1H}-b_{1H})=0$
2.	$n=0$; $m>0$; $K_1=1$	a. $m=1$; $1+K_1(a_{1H}-b_{1H})=0$
		b. $m>1$; $a_{1H}=b_{1H}$
3.	$n<0$; $m=0$; $K_1=1$	a. $n=-1$; $a_1a_0=b_1a_0+K_3$
		b. $n<-1$; $a_1a_0=b_1a_0$

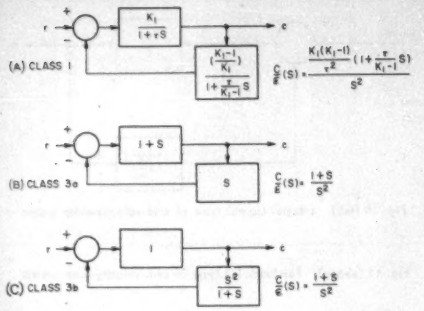
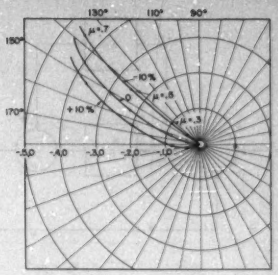


Fig. 12 (left). Examples of other zero-velocity-error types

Fig. 14 (right). Inverse plane loci for variation in forward gain



the system error and not the actuating signal in this case.

Derivation and Classification of Zero-Velocity-Error Systems

This system is a special case of a more general class (use of a lagging doublet in the feedback path) of zero-velocity-error systems described by King.[1] In Appendix I, a technique is presented which, it is felt, derives systems of this type in a more straight forward and simpler manner than that used by King. Both the method in Appendix I and the closed-loop parameter method used by King require the use of trial forms. In Appendix II, a derivation is presented which yields a complete classification of all possible zero-positional error (type-1) and zero-velocity error (type-2) systems having a $G(s)\ H(s)$ form where the $G(s)$ and $H(s)$ are ratios of polynomials. The results are summarized as

$$\text{For a } G(s) = \frac{K_1(1+a_{10}s+a_{20}s^2+\dots)}{s^m(1+b_{10}s+b_{20}s^2+\dots)}$$

$$\text{and an } H(s) = \frac{K_2(1+a_{1H}s+a_{2H}s^2+\dots)}{s^n(1+b_{1H}s+b_{1H}s^2+\dots)}$$

Then the only possible conditions for a type-1 system are those shown in the left-hand column of Table I. The only possible conditions for a type-2 system are those in the left-hand column plus those in the right-hand column in Table I. It is observed that the system under discussion is a special case of class *2a*. The integral control system is a special case of class *2b*. Fig. 12(A), (B), and (C) depicts examples of classes *1*, *3a*, and *3b* respectively. It is noted that the C/E transfer functions for the examples shown are of the same -12, -6 form as that discussed in this paper. The method in Appendix II can be extended to obtain conditions for type-3 and higher systems.

Effects of Parameter Variations

The immediate question upon seeing systems of this type, where a critical set of conditions exists, is: "What happens when the conditions are not exactly met, when the parameters vary?" It will be shown that small parameter variations have very little effect on system stability but, instead of zero-velocity error, a small velocity-lag or velocity-lead error occurs.

In Appendix II, the expression for the steady-state error to a unit step of velocity for a class-*2a* system was shown to be

$$E_{ssp} = \frac{1+K_1(a_{1H}-b_{1H})}{K_1} \qquad (14)$$

In this feedback lag system $a_{1H}=0$ and b_{1H} is ideally set to $1/K_1$, where $K_1=\omega_C$. If b_{1H} should vary so that $b_{1H}=(1/\omega_C)(1+\Delta)$, then

$$E_{ssp} = \frac{1+\omega_C\left(-\frac{1+\Delta}{\omega_C}\right)}{\omega_C} = -\frac{\Delta}{\omega_C} = -\frac{1}{\omega_C/\Delta} \qquad (15)$$

Similarly, if K_1 should vary so that $K_1=\omega_C(1+\Delta)$, then

$$E_{ssp} = \frac{1+\omega_C(1+\Delta)\left(-\frac{1}{\omega_C}\right)}{\omega_C(1+\Delta)} = \frac{-\Delta}{\omega_C(1+\Delta)} \doteq \frac{-1}{\omega_C/\Delta} \qquad (16)$$

Thus, if Δ is negative, a small velocity-lag error results with an effective $K_v=\omega_C/\Delta$. If Δ is positive, a small velocity-lead error results. If Δ is 1 per cent for example, a $K_v=100\omega_C$ will be obtained. An allowable tolerance on Δ can be set according to system requirements. An adjustment should be provided in the system to set in the precise parameter conditions and to reset them from time to

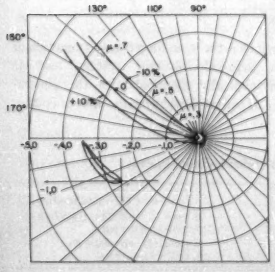


Fig. 13 (left). Inverse plane loci for variation in feedback lag time constant

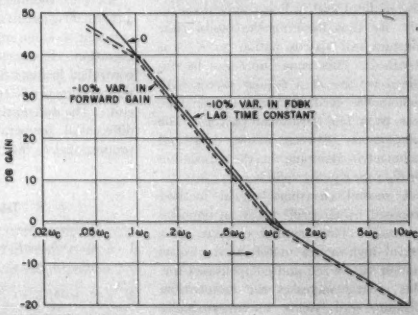


Fig. 15 (right). Db-log ω plot for feedback lag system with parameter variations

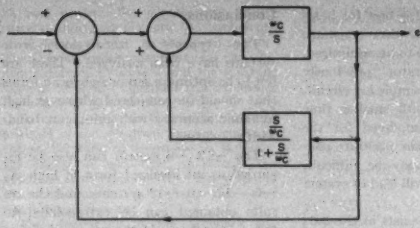


Fig. 16 (above). System equivalent to feedback lag type

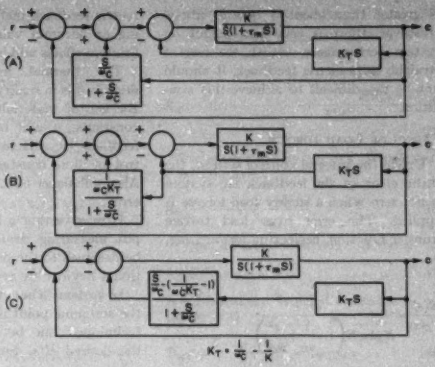


Fig. 17 (right). Tachometric systems equivalent to feedback lag type

time. The use of precision tachometric feedback should make it possible to keep parameter variations well under 1 per cent.

Observation of the characteristic equation with and without parameter variation shows that, for small variations, there is very little effect on stability. Thus, $1+G(s)H(s)=0$ without variation becomes

$$1+\frac{\omega_C}{s}\left(\frac{1}{1+\frac{1}{\omega_C}s}\right)=0$$

$$\text{or } s^2+\omega_C s+\omega_C{}^2=0 \qquad (17)$$

and with variation becomes

$$1+\frac{\omega_C(1+\Delta_1)}{s}\left(\frac{1}{1+\frac{1}{\omega_C}(1+\Delta_2)s}\right)=0$$

$$\text{or } s^2+\frac{\omega_C}{1+\Delta_2}s+\omega_C{}^2\frac{(1+\Delta_1)}{(1+\Delta_2)}=0 \qquad (18)$$

This is further demonstrated in Figs. 13 and 14, where the inverse plane loci of $E/C\ (j\omega)$ for ±10-per-cent variation in the parameters are plotted, and in Fig. 15, where the db-log ω plots of $C/E\ (j\omega)$ for −10-per-cent variation in the parameters are depicted. The reversion to a type-1 −6 −12 −6 system are observed in these plots.

It is interesting to note that the transfer function

$$\frac{C}{E}(s)\cong\frac{\omega_C\left(1+\frac{s}{\omega_C}\right)}{\Delta s\left(-1+\frac{s}{\Delta\omega_C}\right)} \qquad (19)$$

becomes unstable for a positive value of Δ (the inverse transfer function $E/C\ (s)$ becomes nonminimum phase). The physical open-loop $G(s)\ H(s)$ and the closed-loop stability, however, are hardly affected. One peculiar result of this situation is the occurrence of velocity-lead error. The output position leads the input in the steady-state response to a step of velocity.

Effect of a Forward Lag

At this point, it is of interest to refer to Marcy[4] and to pages 229 and 230 by Brown and Campbell[3] where an analysis of a system with a feedback lag is undertaken. Brown and Campbell leave the impression that a feedback lag is undesirable because it has a destabilizing effect. It is unfortunate that the references chose an example where the lag time constant in the forward path was equal to or greater than that in the feedback path. If a case had been studied where the forward lag time constant was much smaller than the feedback lag time constant, a peak M as low as 1.47 would have been attained with an infinite K_p. Use of a lagging doublet in the feedback path instead of a simple lag could achieve M of 1.3 and lower. The advantages of systems of this type would have been realized much sooner.

However, the example does point out the importance of looking at the effect of a forward lag. A forward lag in this context is the lag term in a resultant $G(s)$, not necessarily a lag in the physical forward path of the servo

For a $G(s)=\dfrac{\omega_C}{s(1+\tau s)}$

and an $H(s)=\dfrac{1}{1+\dfrac{s}{\omega_C}}$

the resulting

$$\frac{C}{E}(s)=\frac{\omega_C{}^2\left(1+\frac{s}{\omega_C}\right)}{(1+\omega_C\tau)s^2\left(1+\frac{\tau}{1+\omega_C\tau}s\right)} \qquad (20)$$

If $\omega_C\tau<<1$ or $\tau<<1/\omega_C$, then

$$\frac{C}{E}(s)\approx\frac{\omega_C{}^2\left(1+\frac{s}{\omega_C}\right)}{s^2(1+\tau s)} \qquad (21)$$

Thus, the effect of a small lag in $G(s)$ is the same in this system as in the integral control system; the db-log ω plot becomes −12, −6, −12 and, as long as the breakpoint caused by the forward lag

Table II. Possible Combinations for m and n

Combination	Steady-State Error	Conditions for Zero Error
$m=0;\ n=0$	$E_{ssd}=\dfrac{1+K_1K_2-K_3}{1+K_1K_2}$	$1+K_1K_2-K_3=0$
$m>0;\ n=0$	$E_{ssd}=\dfrac{K_1K_2-K_3}{K_1K_2}$	$K_3=1$
$m=0;\ n<0$	$E_{ssd}=1-K_3$	$K_3=1$
$m<0;\ n\leq 0$	$E_{ssd}=1$	none
$m\geq 0;\ n>0$	$E_{ssd}=\dfrac{K_1K_2}{K_1K_2}=1$	none
$m>0;\ n<0$	$E_{ssd}=\infty$	none
$m<0;\ n>0;\ \lvert m\rvert=\lvert n\rvert$	$E_{ssd}=\dfrac{1+K_1K_2}{1+K_1K_2}=1$	none
$m<0;\ n>0;\ \lvert m\rvert>\lvert n\rvert$	$E_{ssd}=1$	none
$m<0;\ n>0;\ \lvert m\rvert<\lvert n\rvert$	$E_{ssd}=\dfrac{K_1K_2}{K_1K_2}=1$	none

is greater than about 3 octaves above crossover, there is very little effect on system performance. With the use of straight tachometric feedback, it should not be too difficult to achieve this condition.

Effect of Load Torque

Unlike the integral control system, the static error of the feedback lag system is not zero when a steady load torque is applied. The error over load torque transfer function, neglecting forward lag, is

$$\frac{E(s)}{T_L(s)} = \frac{1+\frac{s}{\omega_C}}{K_A k_m N\left(1+\frac{s}{\omega_C}+\frac{s^2}{\omega_C^2}\right)} \qquad (22)$$

The static error to a steady load torque is, like in most servomechanisms, given by

$$E_{ss} = \frac{T_L}{K_A k_m N} \qquad (23)$$

where

T_L = effective static torque at load shaft
K_A = the gain from summing point to motor

This could be reasoned physically since, in the steady state, the feedback is unity, and there is no integration up to the motor torque. Thus, it is observed that zero-velocity-error systems do not necessarily yield zero torque error.

Methods of Obtaining $1/(1+s/\omega_C)$

Electronic, electromechanical, and thermal means may also be employed for obtaining the feedback lag. In d-c systems, a simple resistance-capacitance passive circuit can be used. In a-c systems, a demodulator d-c lag circuit-modulator combination could be employed. In such a case the lag circuit will help to attenuate demodulator ripple; however, gain variation, unbalances, wave form, and modulator noise problems will exist. Resistance-inductance-capacitance resonant circuits can also be employed for a-c systems. To minimize the size of choke required, Q multiplier[5] circuits can be used in conjunction. In these circuits, the required tight control of Q may be difficult to achieve. The usual frequency drift problems exist. Active resistance-capacitance circuits which behave like reistance-inductance capacitance circuits[6] can also be used in a-c systems.

Any of the electromechanical methods previously described for obtaining K/s can be utilized to obtain a simple lag $1/(1+s/K)$ by use of unity negative feedback around the device. The tachometric feedback method is best for holding an accurate value of time constant. Stiction effects will have to be minimized.

The thermal integrator previously described is in reality a simple lag circuit. By use of materials with smaller time constants, it can be employed for the feedback lag. Here again, accurate control of time constant may be difficult. Also, unbalance effects will lead to system error.

It may be that a lag exists in the output measuring device. If this lag can be designed to the proper value, no additional devices are required.

In systems where the feedback path to the summing point is unattainable, other techniques can be employed to yield the desired $H(s)$ function. Thus

$$\frac{1}{1+\frac{s}{\omega_C}} = 1 - \frac{\frac{s}{\omega_C}}{1+\frac{s}{\omega_C}}$$

which can be obtained by the block diagram in Fig. 16. The new transfer function can be obtained by most of the methods described in the foregoing. If tachometric feedback is imployed to obtain the forward path, then the feedback paths can be combined, as shown in Fig. 17(A), (B), and (C). It is interesting to point out that problem 10.7 by Chestnut and Mayer[7] is a system of this form. It can also be observed that tachometric feedback through a high-pass filter is an approximation to the desired transfer function, as can be expected. As the tachometric loop gain approaches infinity, $\omega_C K_T$ approaches unity and the network of Fig. 17(C) approaches a high-pass filter.

Comparison of Systems

The integral control system is much less sensitive to parameter variations than the feedback lag system. It also reduces to zero the static error caused by load torque while the feedback lag system does not.

On the other hand, especially for d-c systems, the circuits required for feedback lag are simpler than those required for integral control. The feedback lag system can also take advantage of any lag that might exist in the output measuring device. The feedback lag should also attenuate any noise (such as potentiometer noise) that might be generated in the output measuring device.

The choice of which system to employ depends on the relative importance of these and other factors to the particular application at hand.

Conclusions

Two types of zero-velocity-error systems have been analyzed. These are felt to be optimum servomechanism forms that should be considered whenever high dynamic accuracy with minimum bandwidth is desired.

The -12 -6 system can also be regarded as an idealized form of high K, -6 -12 -6 -12 systems, and the results obtained can be extrapolated to those systems. Methods of deriving zero-forced-error systems and a classification of possible zero-velocity-error systems were presented which should be very useful for any further work along these lines. Lastly, the technique of using the response to an acceleration-velocity input function as a measure of dynamic accuracy is felt to be most meaningful and can be extended to all servomechanism types.

Appendix I. Derivation of Zero-Forced-Error Systems

A general technique of deriving zero-forced-error systems is as follows. Since

$$\frac{C}{R}(s) = \frac{G(s)}{1+G(s)H(s)} = \frac{\frac{C}{E}(s)}{1+\frac{C}{E}(s)} \qquad (24)$$

Given a $G(s)$ and an $H(s)$, then solving equation 1

$$\frac{C}{E}(s) = \frac{G(s)}{1+G(s)H(s)-G(s)} \qquad (25)$$

$$\frac{C}{E}(s) = \frac{A(s)}{s^n B(s)} \qquad 4$$

where

$A(s)$ and $B(s)$ = polynomials in s
n = a type number for a system which directly determines the steady-state forced error

If $n=0$, there is a steady-state error to a step of position (regulator). If $n=1$, the steady-state positional error is zero and a velocity lag to a step of velocity exists. If $n=2$, there is zero-velocity error as well as zero-positional error and an acceleration lag exists, etc.

Thus for zero-velocity-error systems, the lowest order exponent appearing in the denominator of equation 25 must be 2. For zero-acceleration-error systems, the lowest exponent must be 3, etc.

The problem to be solved is the following: Given certain parts of a feedback control system, what additional transfer functions are required to yield a type-n system. For example a $G_n(s)$, where $G(s)=G_1(s)G_n(s)$, and an $H(s)$ may be given and a $G_1(s)$ is desired that will yield a type-n system. Or a $G(s)$ may be given and an $H(s)$ is desired that will yield a type-n system.

The method of solution is to let the unknown $F(s)$ equal some general trial form; substitute this form into equation

25 and expand the denominator; then successively equate to zero the coefficients of s^0, s^1, ..., s^{m-1} in the denominator. The lowest order exponent in the denominator will then be n and the n simultaneous coefficient equations will determine the required parameters of $F(s)$.

To establish its practical utility, further analysis of the resultant system must, of course, be made to determine its stability, the effects of parameter variations, the synthesis of the required function, response to other inputs, peak error, and other usual design factors.

The derivation of the feedback lag zero-velocity-error (type-2) system will serve as an illustration of the technique.

Given: $G(s)=\omega_C/s$. Find: an $H(s)$ to yield a type-2 system. Let

$$H(s)=K\frac{N(s)}{D(s)}=K\frac{(1+a_1s+a_2s^2+\dots)}{(1+b_1s+b_2s^2+\dots)}$$

Then

$$\frac{C}{E}(s)=\frac{\dfrac{\omega_C}{s}}{1+\dfrac{\omega_C}{s}K\dfrac{N(s)}{D(s)}-\dfrac{\omega_C}{s}}=\frac{\omega_CD(s)}{sD(s)+K\omega_CN(s)-\omega_CD(s)}$$

Equating coefficients of $s^0=0$

$$K\omega_C-\omega_C=0$$

Coefficients of $s^1=0$

$$1+K\omega_Ca_1-\omega_Cb_1=0$$

Solving these simultaneous equations yields $K=1$ and $b_1-a_1=1/\omega_C$. Any $H(s)$ of the trial form that satisfies these relationships will yield at least a type-2 system. The simplest such $H(s)$ is the simple lag $1/(1+1/\omega_Cs)$.

Appendix II. Derivation of Possible Zero-Velocity-Error Systems

Let

$$G(s)=\frac{K_1N_G(s)}{s^mD_G(s)}$$

and

$$H(s)=\frac{K_2N_H(s)}{s^nD_H(s)} \tag{26}$$

where

$$N_G(s)=1+a_{1G}s+a_{2G}s^2+\dots$$
$$D_G(s)=1+b_{1G}s+b_{2G}s^2+\dots$$
$$N_H(s)=1+a_{1H}s+a_{2H}s^2+\dots$$
$$D_H(s)=1+b_{1H}s+b_{2H}s^2+\dots$$

Also, the following relationships will be used

$$\lim_{s\to0}N_G(s)=\lim_{s\to0}D_G(s)=\lim_{s\to0}N_H(s)=\lim_{s\to0}D_H(s)=1 \tag{27}$$

$$\lim_{s\to0}\frac{d}{ds}N_G(s)=a_{1G}\lim_{s\to0}\frac{d}{ds}N_H(s)=a_{1H}$$

$$\lim_{s\to0}\frac{d}{ds}D_G(s)=b_{1G}\lim_{s\to0}\frac{d}{ds}D_H(s)=b_{1H} \tag{28}$$

$$\frac{E(s)}{R(s)}=\frac{1+G(s)H(s)-G(s)}{1+G(s)H(s)}=\frac{1+\dfrac{K_1K_2N_G(s)N_H(s)}{s^{m+n}D_G(s)D_H(s)}-\dfrac{K_1N_G(s)}{s^mD_G(s)}}{1+\dfrac{K_1K_2N_G(s)N_H(s)}{s^{m+n}D_G(s)D_H(s)}} \tag{29}$$

By the final value theorem

$$E_{ss}=\lim_{s\to0}sE(s)=\lim_{s\to0}sR(s)\times\left(\frac{s^{m+n}D_G(s)D_H(s)+K_1K_2N_G(s)N_H(s)-s^nK_1N_G(s)D_H(s)}{s^{m+n}D_G(s)D_H(s)+K_1K_2N_G(s)N_H(s)}\right) \tag{30}$$

If $R(s)=1/s$, unit step of displacement

$$E_{ssd}=\lim_{s\to0}\times\left(\frac{s^{m+n}D_G(s)D_H(s)+K_1K_2N_G(s)N_H(s)-s^nK_1N_G(s)D_H(s)}{s^{m+n}D_G(s)D_H(s)+K_1K_2N_G(s)N_H(s)}\right) \tag{31}$$

The combinations of m and n which are possible are shown in Table II. For each combination, E_{ssd} is obtained with the use of the foregoing relationships and the conditions for $E_{ss}=0$ are determined.

If $R(s)=1/s^2$, unit step of velocity, then

$$E_{ssv}=\lim_{s\to0}\frac{E_{ssd}}{s} \tag{32}$$

Thus, for the steady-state error to a step of velocity to be finite, the steady-state error to a step of displacement must be zero. This is well known and can also be ascertained from the fact that the response to a step of velocity is the integral of the response to a step of displacement. Therefore, to determine possible zero-velocity-error systems, only the first three combinations need be considered. The conditions for zero E_{ssd} are also necessary for zero E_{ssv}, but the following additional conditions are also necessary. By L'Hospital's rule

$$E_{ssv}=\lim_{s\to0}\times\frac{\dfrac{d}{ds}\left(s^{m+n}D_G(s)D_H(s)+K_1K_2N_GN_H(s)-s^nK_1N_G(s)D_H(s)\right)}{\dfrac{d}{ds}\left\{s[s^{m+n}D_G(s)D_H(s)+K_1K_2N_G(s)N_H(s)]\right\}} \tag{33}$$

For class 1, $m=0$; $n=0$. Differentiating and applying the previously stated limit relationships

$$E_{ssv}=\frac{b_{1H}+b_{1G}+K_1K_2a_{1H}+K_1K_2a_{1G}-K_1b_{1H}-K_1a_{1G}}{1+K_1K_2} \tag{34}$$

Applying the relation $1+K_1K_2-K_1=0$ which must be satisfied for E_{ssd} to equal zero

$$E_{ssv}=\frac{b_{1G}-a_{1G}+(K_1-1)(a_{1H}-b_{1H})}{1+K_1K_2} \tag{35}$$

Thus for $E_{ssv}=0$ for this class

$$b_{1H}-a_{1G}+(K_1-1)(a_{1H}-b_{1H})=0 \tag{36}$$

is an additional required condition.

For class 2, $m>0$; $n=0$

$$E_{ssv}=\lim_{s\to0}\frac{ms^{m-1}+K_1K_2a_{1H}+K_1K_2a_{1G}-K_1b_{1H}-K_1a_{1G}}{K_1K_2} \tag{37}$$

For $m=1$, and also applying the necessary condition $K_1=1$

$$E_{ssv}=\frac{1+K_1a_{1H}-K_1b_{1H}}{K_1} \tag{38}$$

Thus for $E_{ssv}=0$ for this subclass

$$1+K_1(a_{1H}-b_{1H})=0 \tag{39}$$

is an additional required condition.

For $m>1$

$$E_{ssv}=\frac{K_1(a_{1H}-b_{1H})}{K_1}=a_{1H}-b_{1H} \tag{40}$$

Thus for $E_{ssv}=0$ for this subclass

$$a_{1H}=b_{1H} \tag{41}$$

is an additional required condition.

For class 3, $m=0$; $n<0$

$$E_{ssv}=\lim_{s\to0}\frac{b_{1H}+b_{1G}-ns^{-n-1}K_1K_2-K_1b_{1H}-K_1a_{1G}}{1} \tag{42}$$

For $n=-1$, and also applying the necessary condition $K_1=1$

$$E_{ssv}=b_{1G}+K_2-a_{1G} \tag{43}$$

Thus for $E_{ssv}=0$ for this subclass

$$a_{1G}-b_{1G}=K_2 \tag{44}$$

is an additional required condition.

For $n<-1$

$$E_{ssv}=b_{1G}-a_{1G} \tag{45}$$

Thus for $E_{ssv}=0$ for this subclass

$$a_{1G}=b_{1G} \tag{46}$$

is an additional required condition. These results are summarized in Table I.

References

1. Reduction of Forced Errors in Closed Loop Systems, L. H. King. *Proceedings*, Institute of Radio Engineers, New York, N. Y., Aug. 1953, pp. 1037-42.

2. Principles of Servomechanisms (book), G. S. Brown, D. P. Campbell. John Wiley and Sons, Inc., New York, N. Y., 1948.

3. Integral Controller for Use in Carrier-Type Servomechanisms, R. L. Cosgriff. *AIEE Transactions*, vol. 69, 1950, pp. 1379-83.

4. Parallel Circuits in Servomechanisms, H. Tyler Marcy. *Ibid.*, vol. 65, Aug.-Sept. 1946, pp. 521-29.

5. Simplified Q Multiplier, H. E. Harris. *Electronics*, New York, N. Y., May 1951, p. 130.

6. A Practical Method of Designing R. C-Active Filters, R. P. Sallen, E. L. Key. *Transactions on Circuit Theory*, Institute of Radio Engineers, New York, N. Y., vol. CT-2, no. 1, Mar. 1955.

7. Servomechanisms and Regulating System Design (book), H. Chestnut, R. W. Mayer. John Wiley and Sons, Inc., New York, N. Y., 1951, vol. 1.

Inert Gas-Shielded Welding Arc Behavior and Metal Transfer Characteristics

G. M. SKINNER
ASSOCIATE MEMBER AIEE

D. M. YENNI
NONMEMBER AIEE

Synopsis: An investigation of inert gas-shielded consumable electrode arc behavior and metal transfer characteristics was undertaken on aluminum and steel in flat, vertical, and overhead welding positions with the use of argon, argon+5-per-cent(%) oxygen, helium and helium+5% oxygen shielding gases. Arc and metal transfer behavior were observed with high-speed motion pictures with the use of intense supplementary illumination to facilitate photography. Electrical behavior of the arc was examined during production of a considerable number of weld beads by recording oscilloscope traces of arc current or voltage on the high-speed pictures of the arc. Thus, relationships were observed between current transients and arc physical behavior. Arcs shielded with argon or a mixture of argon+5% oxygen, known commercially as sigma grade argon, were more stable and operated with less spatter than similar arcs in helium or helium+5% oxygen. Although addition of oxygen to argon was desirable, addition of oxygen to helium shielding gas did not improve arc behavior or metal transfer. Numerous large transients occurred in current and voltage of reverse polarity (RP) arcs in helium and helium+oxygen shielding gas. The transients or disturbances were frequently associated with drop severance from the filler wire and with formation of undesirable spatter. Electrical disturbances were not present in similar arcs in argon or argon+oxygen shielding gases.

DURING the past few years, considerable information from various sources has appeared about welding arc behavior with the use of consumable and nonconsumable electrodes in several shielding gases on a variety of base metals. Research to provide information has been in progress at the Linde Air Products Company Laboratories continuously in connection with development of Heliarc* and sigma welding. This discussion is the third in a series of papers dealing with the behavior of inert gas-shielded welding arcs.

The first paper discussed voltage distribution in the arc core, heat transfer measurements and metal transfer behavior in steel, and aluminum welding with argon shielding.[1] More detailed information on metal transfer was given in a second paper in which the mechanism of metal transfer in pure argon and in argon +5% oxygen mixtures, with the use of straight and reverse polarity on steel, stainless steel, and aluminum, was discussed.[2] Information on metal transfer appeared recently in an article describing research with coated electrodes.[3]

Investigation of arc and metal transfer behavior has now been extended to arcs shielded with argon, sigma argon, helium, and helium +5% oxygen mixtures operating on steel and on aluminum in flat, vertical, and overhead welding. However, only a limited amount of data is presented for aluminum welding because the sigma-process behavior on this base metal has rapidly become well known. Currently, a large amount of research attention is being given to steel welding where porosity and other problems still exist. It is recognized that position welding with inert gas shielding on steel is still in the development stage. The information in this paper is, therefore, primarily valuable in experimental work with the process.

The research described herein was undertaken primarily to learn the influence of shielding-gas composition on arc stability, spatter formation, and efficiency of metal transfer. All of these characteristics must be considered in establishing satisfactory welding techniques. Before presenting the results, a brief description is given of innovations in the experimental methods and photography.

Experimental Methods and Photography

Experience in the past showed that high-speed motion pictures provide a desirable method of studying arc and metal transfer behavior and that oscilloscope recordings of arc voltage and current give considerable information about arc stability. The photographic procedure was a modification and improvement of the method described in detail by Gillette and Breymeier.[1] The simultaneous trace and arc pictures were taken with a camera equipped with an auxiliary optical system for recording oscilloscope traces on the back of the film. This process unfortunately requires the use of negative instead of black and white Cine film or Kodachrome. For this reason, many of the photographs were taken without the synchronized traces.

The high intensity and large variations in intensity of light from the welding arc region created difficult exposure problems when efforts were made to secure clear high-speed pictures of metal transfer. A good solution was the use of supplementary illumination arranged to show globule transfer in silhouette by directing an auxiliary high-intensity light beam directly into the camera lens through the arc, as indicated in Fig. 1. The arc lamp employed for this purpose was an outdoor motion-picture type operated at 7 kw. As shown in Fig. 1, flat weld beads were made by mounting the workpiece on the Linde* *CM-37* carriage equipped with an electronic governor to obtain uniform movement. Vertical position welds were made on ratating disks. Linear travel speeds of about 20 inches per minute were used for flat welding and 10 inches per minute for position welding. A sigma welding machine was provided with power from a rectifier-type, conventional d-c source.

Aluminum Welding

Experimental data showing the effect of shielding-gas composition on art behavior are summarized in Table I. Changes in shielding-gas composition greatly affected arc stability and spatter in aluminum welding with RP arc current. Investigation of straight polarity (SP) current performance on aluminum was not included in this phase of the research. Arcs in argon and sigma argon shielding were stable in shape and position. Spatter was negligible. The range of changes in arc length and arc potential was relatively low. In con-

*Registered trade-mark of the Union Carbide and Carbon Corporation.

Paper 56-112, recommended by the AIEE Electric Welding Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Winter General Meeting, New York, N. Y., January 30-February 3, 1956. Manuscript submitted April 11, 1955; made available for printing November 21, 1955.

G. M. SKINNER and D. M. YENNI are with the Linde Air Products Company, Tonawanda, N. Y.

Assistance and suggestions from members of the Linde Air Products Laboratory staff contributed to the investigation described in this paper. Particular recognition is given to the work of W. J. Semerau, E. J. Smolarek, and K. L. Thomas who are associated with photography, electronics, and welding techniques respectively.

Fig. 1. Arrangement of experimental equipment

trast, arcs in helium and helium + 5% oxygen shielding gases moved rapidly and erratically on the base metal causing large amounts of spatter. Arc length, as used herein, signifies the distance between the electrode tip and the workpiece surface. Crater depth was not measured.

Changes in shielding-gas composition did not have much influence on the velocity or size of droplets of molten aluminum metal transferred from the filler wire to the work. Globule diameters varied from 0.02 to 0.10 inch. Globule velocity was approximately 10 feet per second in each of the shielded gases investigated.

The range of arc potentials shown in Table I for the flat position was measured by oscilloscope traces taken simultaneously with photographs yielding data on variations in arc length. Fairly good correlation exists between percentage changes in arc potential and arc length. Furthermore, arcs operating with uniform arc length were relatively stable and free from spatter.

Steel Welding—Flat Position

The influence of shielding-gas composition on arc and metal transfer characteristics on steel in the flat welding position is indicated by the experimental data summarized in Table II. Arc behavior, metal transfer characteristics, and spatter formation in RP, argon, and sigma argon shielded arcs were satisfactory. The arcs were stable in length and position on the base metal. Transfer of molten globules from the welding wire to the puddle took place in uniform size, with relatively small diameter drops moving at high velocity. The formation of spatter was negligible. It is immediately evident that the SP arc is not as desirable as the RP arc because of: 1. large variations in arc length; 2. low velocity transfer of relatively large molten metal globules across the arc; 3. poor arc stability; and 4. heavy spatter formation. However, the chief effects of adding oxygen to argon shielding were observed in SP welding where the oxygen addition reduced the variation in arc length and the diameter of molten droplets. A large reduction in spatter was also observed.

Addition of oxygen to helium did not improve arc or metal transfer characteristics in either RP or SP arcs. In fact, the variation in arc length experienced during RP welding was increased. No change was observed in arc stability or spatter.

Steel Welding—Vertical Position

Data in Table II on steel in the vertical position indicate good arc stability; generally satisfactory metal transfer behavior and low spatter formation were characteristic of RP arcs shielded with argon and sigma argon. Helium-shielded, RP arc properties were not as desirable as those of the argon-shielded arcs. The behavior of helium +5% oxygen-shielded arcs was still less desirable. Advantages of the RP arcs in argon-rich gases are: 1. high percentages of metal transferred from the filler wire to the welding puddle; 2. small, rapidly moving globules; 3. stable arc length; and 4. negligible spatter. All or part of large-diameter, slowly moving globules did not reach the weld puddle in the vertical position because of relatively large acceleration by gravity normal to the drop transit direction. The actual velocity of the molten metal drops was calculated from a number of the high-speed pictures with the procedure indicated in Fig. 2. Velocities ranged from 0.1 to 10 feet per second.

Data plotted in Fig. 2 show the velocity remained constant for most of the droplet passage across the 0.22-inch-long arc. However, some acceleration took place as the globule approached the base plate. In the last 0.075 inch of travel the velocity increased from 4.2 to 6.3 feet per second. The total transit time for the specific example used in Fig. 2 is about 0.004 second. In this length of time the

Table I. Arc and Metal Transfer Characteristics on Aluminum, Flat and Vertical Welding Position

Shielding Gas	Welding Position	Arc Potential Range, Volts	Arc Length Range, Inches	Globule Diameter Range, Inches	Arc Stability	Spatter
Argon	flat	21.5–23.7	0.23–0.27	0.05–0.09	good	none
Argon +5% oxygen	flat	20.8–23.2	0.20–0.23	0.08–0.08	good	none
Helium	flat	22.8–26.2	0.15–0.19	0.04–0.08	poor	heavy
Helium +5% oxygen	flat	22.2–30.0	0.06–0.09	0.04–0.10	poor	heavy
Argon	vertical	23	0.37–0.47	0.05–0.05	good	none
Argon +5% oxygen	vertical	23	0.43–0.48	0.05–0.05	good	none
Helium	vertical	28	0.11–0.15	0.03–0.06	poor	heavy
Helium +5% oxygen	vertical	28	0.08–0.15	0.02–0.06	poor	heavy

Arc current: d-c RP, 210 amperes flat, 180 amperes vertical position.
Globule velocity: approximately 10 feet per second for all conditions.
Welding wire: Oxweld* no. 23, 1/16-inch diameter.
Shielding-gas consumption: 50 cubic feet per hour for argon, 100 cubic feet per hour for helium.
Welding-speed: flat—20 inches per minute; vertical—10 inches per minute.
Base plate: 1/4-inch 2S aluminum.

* Registered trade-mark of the Union Carbide and Carbon Corporation.

Table II. Arc and Metal Transfer Characteristics on Steel, Flat Welding Position

Shielding Gas	Arc Length Range, Inches	Globule Diameter Range, Inches	Approximate Globule Velocity	Arc Stability	Spatter
RP					
Argon	0.16–0.20	0.03–0.05	high	good	none
Argon + 5% oxygen	0.16–0.20	0.03–0.05	high	good	none
Helium	0.06–0.20	0.02–0.08	medium	poor	heavy
Helium + 5% oxygen	very erratic	0.02–0.08	low	poor	heavy
SP					
Argon	erratic	0.13–0.20	low	poor	heavy
Argon + 5% oxygen	0.19–0.37	0.06–0.14	medium	poor-fair	medium
Helium	erratic	0.05–0.27	low	poor	very heavy
Helium + 5% oxygen	0.06–0.22	0.05–0.16	low	poor	very heavy

Arc current: 280 amperes RP, 340 amperes SP.
Arc potential: argon—28 volts, helium—35 volts.
Welding wire: Oxweld no. 32 *CMS*, 1/16-inch diameter.
Shielding-gas consumption: 50 cubic feet per hour for argon, 100 cubic feet per hour for helium.
Welding speed: 20 inches per minute on 1/4-inch mild steel plate.

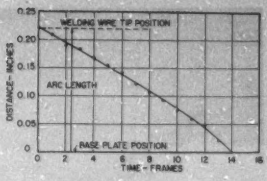


Fig. 2. Molten metal transit velocity, vertical welding position on steel

Shielding gas: 75 cubic feet per hour for argon
Arc current: 195 amperes d-c RP
Arc potential: 25 volts
Welding wire: Oxweld no. 32 CMS, 3/64-inch diameter
Camera speed: 3,225 frames per second, one frame = 0.00031 second
Average drop velocity: 0.22/(12×14×0.00031) = 4.2 feet per second
Maximum droplet velocity: (4.2×0.33)/0.22 = 6.2 feet per second

distance travelled by the droplet in the vertical direction because of gravitational forces is about 0.003 inch, a negligible distance compared to the 0.22-inch horizontal arc length. Gravitational forces become predominant at the lower globule velocities.

The beneficial effect of adding oxygen to argon in SP welding is demonstrated by the data in Table III, which indicates a reduction in molten metal drop size and an increase in metal transfer efficiency. Improved arc stability and greatly increased droplet velocity also resulted from addition of oxygen to argon in SP welding. The addition of oxygen to helium had the effect of reducing variations in arc length and globule diameter but did not improve arc stability, spatter, or metal transfer characteristics.

These observations are in agreement with the well-known influence of oxygen additions to argon for shielding SP and RP arcs operating at lower current densities. The oxygen additions result in improved stability, reduced spatter, increased metal globule velocity, and reduced metal droplet diameter.

It will be observed that large shielding-gas consumptions were used in obtaining the data presented in Table III to avoid smoke interference with photography. However, the smoke made the shielding flow patterns visible, thereby indicating why large flow rates of light gases are necessary in vertical welding. This welding geometry located the entire hot-metal area directly below the arc. The mass of hot metal caused much stronger convection atmosphere currents at the arc area in the vertical welding position than in conventional flat and in overhead positions. Furthermore, the shielding gas was not supplied with an initial downward velocity component.

The correlation between irregularities in arc length, arc stability, and spatter formation previously observed in aluminum welding also exists in steel welding. This behavior and the ability of the constant potential power source to correct changes in arc length suggest the use of constant potential power as a means of reducing arc instability and spatter.

Oscilloscopic recordings of arc voltage and current confirmed the results of high-speed photography showing arcs in argon or sigma argon were more stable than those in helium or helium +5% oxygen shielding gas. Data summarized in Table IV show disturbances (rapid changes or transients) in arc potential and current occurred frequently in arcs shielded with helium or helium +5% oxygen gas mixtures. The transients were observed more frequently in the helium + oxygen-shielded arcs than in the helium-shielded arcs and were entirely absent in arcs shielded with argon or sigma argon. Typical, simultaneously recorded traces and frames of high-speed motion pictures are reproduced in Fig. 3.

A transient may or may not be associated with severence of a relatively large molten droplet from the filler wire. Examples of both kinds of transients are included in Fig. 3. Furthermore, droplets are frequently severed without causing an electrical disturbance. It is interesting to observe that the transients appeared at a rate of about five per second in helium-shielded arcs and that they generally involved an increase in arc potential and a reduction in arc current. Some of the irregularities are damped oscillations with a length varying from 10 to 100 microseconds. The maximum change from stable operating conditions observed was about 50% of the arc current and 75% of the arc potential. Rapid-duration short circuits also existed

Table III. Arc and Metal Transfer Characteristics on Steel, Vertical Position

Shielding Gas	Arc Length Range, Inches	Globule Diameter Range, Inches	Approximate Globule Velocity	Metal Transfer Effect	Arc Stability	Spatter
RP						
Argon	0.13–0.17	0.02–0.04	high	good	good	none
Argon + 5% oxygen	0.13–0.16	0.02–0.04	high	good	good	none
Helium	0.06–0.26	0.04–0.10	medium	medium	medium	medium
Helium + 5% oxygen	0.08–0.36	0.03–0.09	medium	low	medium	heavy
SP						
Argon	0.01–0.14	0.10–0.14	low	low	poor	heavy
Argon + 5% oxygen	0.02–0.19	0.03–0.05	high	good	medium	heavy
Helium	0.04–0.12	0.06–0.18	low	low	medium	heavy
Helium + 5% oxygen	0.06–0.45	0.03–0.11	low	low	poor	heavy

Arc voltage: approximately 25 in argon and 33 in helium shielding gases.
Arc current: 200 amperes RP, 250 amperes SP.
Welding wire: Oxweld no. 32 *CMS*, 3/64-inch diameter.
Shielding-gas consumption: 75 cubic feet per hour for argon, 150 cubic feet per hour for helium.
High shielding-gas flow rates to avoid smoke interference with photography.
Welding speed: 20 inches per minute on 3/8-inch mild steel plate.

Table IV. Arc Electrical Behavior on Steel, Vertical Position

Shielding Gas	Arc Length Range,* Inches	Arc Potential			Arc Current		
			Transients			Transients	
		Voltage	Number	Voltage	Amperes	Number	Amperes
Argon	0.11–0.20	22.5–28.1	none		177–203	none	
Argon+5% oxygen	0.27–0.34	25.7–31.5	none		172–218	none	
Helium	0.06–0.14	23.7–30.7	.5 per second	+22	170–220	.3 per second	+20/−95
Helium+5% oxygen	0.09–0.19	27.9–38.5	.6 per second	−17/+26	168–222	.5 per second	+30/−94

Welding wire: Oxweld no. 32 *CMS*, 3/64-inch diameter.
Welding speed: 10 inches per minute on 3/8-inch mild steel plate.
RP arc current.

* Arc length for potential measurements.

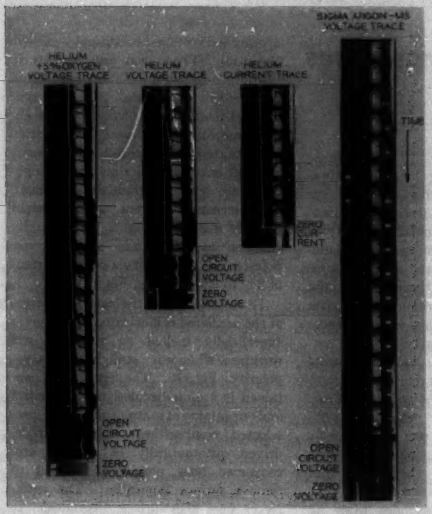


Fig. 3. Simultaneous oscilloscope traces and high-speed motion pictures of sigma arcs on steel, vertical welding position

Arc current: 195 amperes d-c RP
Welding wire: Oxweld no. 32 CMS, 3/64-inch diameter
Welding speed: 10 inches per minute

Table V. Arc and Metal Transfer Characteristics on Steel, Overhead Position

Shielding Gas	Arc Length Range, Inches	Globule Diameter Range, Inches	Approx. Globule Velocity	Metal Transfer Efficiency	Arc Stability	Spatter
RP						
Argon	0.16–0.22	0.01–0.03	high	good	good	none
Argon+5% oxygen	0.02–0.12	0.01–0.03	high	good	good	none
Helium	0.02–0.08	0.04–0.12	medium	medium	medium	medium
Helium+5% oxygen	0.06–0.16	0.03–0.11	low	low	poor	heavy
SP						
Argon	0.02–0.06	0.11–0.13	low	low	poor	heavy
Argon+5% oxygen	0.02–0.13	0.04–0.06	high	good	medium	heavy
Helium	0.01–0.05	0.02–0.18	low	low	poor	heavy
Helium+5% oxygen	0.02–0.06	0.07–0.11	medium	low	poor	heavy

Arc voltage: 25 in argon and 34 in helium shielding gases.
Arc current: 200 amperes RP, 250 amperes SP.
Welding wire: Oxweld no. 32 *CMS*, 3/64-inch diameter.
Shielding-gas consumption: 50 cubic feet per hour for argon, 100 cubic feet per hour for helium.
Welding speed: 10 inches per minute on 1/4-inch mild steel plate.

in the helium-shielded arcs for intervals as long as 0.001 second. During the short circuits, the arc voltage dropped to zero and the arc current increased appreciably.

The voltage trace for the helium +5% oxygen-shielded arc in Fig. 3 shows an interesting and irregular increase in art voltage following transfer of a large droplet of molten metal. Since the voltage increase is obviously not caused by an increase in arc length, it may be a result of either an unusually high concentration of metal vapor in the arc atmosphere from vaporization of the droplet or of a change in arc atmosphere composition by gases released from the superheated metal globule (CO, CO_2 or others). Action of this kind is probably responsible for some of the arc instability and excessive spatter encountered with helium and helium-rich gas shielding.

Steel Welding—Overhead Position

Data summarized in Table V show RP metal transfer characteristics and arc stability are good in sigma and pure argon, fair in helium, and poor in helium +5% oxygen shielding-gas mixtures. High rates of metal transfer from the filler wire to an overhead base plate appear to be associated with relatively small-diameter rapidly moving metal globules. The addition of oxygen to helium was undesirable because of reduced molten metal velocity, increased spatter, and reduced arc stability

SP welding performance was not as good as that of RP with any of the shielding-gas mixtures investigated because of poor arc stability, heavy spatter, and low metal transfer efficiency from the wire to the overhead welding puddle. However, as in other welding positions, the effect of oxygen additions was greatest in SP welding. A large reduction was secured in average drop size and a greater than order-of-magnitude increase in droplet velocity was obtained by adding 5% oxygen to argon. A similar oxygen addition to helium seemed to increase globule velocity somewhat (in SP welding), but had little influence on arc stability or spatter.

Recapitulation

Examination of oscilloscope traces and high-speed motion pictures leads to the following comments regarding arc and metal transfer behavior of inert gas-shielded consumable electrode arcs on aluminum and steel.

1. RP arcs shielded with argon and sigma argon had greater stability and operated with less spatter than similar arcs in helium or helium+5% oxygen mixtures on aluminum in the flat and vertical position and on steel in all welding positions. (Overhead welding on aluminum was not investigated.)

2. Addition of oxygen to argon made large improvements in straight polarity performance on steel by reducing molten globule diameter, increasing molten globule velocity, increasing arc stability, reducing spatter, and increasing the amount of metal transferred from the welding wire to the puddle in position welding.

3. Addition of oxygen to helium generally did not improve arc or metal transfer behavior. The effect of oxygen additions to helium was actually undesirable in RP operation under certain specific operating conditions because of impaired metal transfer efficiency, decreased arc stability, and increased spatter.

4. Numerous large transients were observed in arc current and voltage of RP arcs in helium and helium oxygen-shielded gas mixtures. Droplet severance from the welding wire caused some of the transients. The transients are associated with formation of heavy spatter and were not detected in RP arcs shielded with argon or sigma argon. All SP arc oscilloscope traces contained current and voltage transients.

5. The range of arc length variations during operation was much greater in RP arcs in helium-base shielding gas than in argon.

6. Arc stability, metal transfer characteristics, and spatter formation were generally more satisfactory in RP arcs than in SP arcs.

References

1. Some Research Techniques for Studying Arcs in Inert Gases. R. H. Gillette, R. T. Breymeier. *Welding Journal*, New York, N. Y., Mar. 1951.

2. Metal Transfer in Sigma Welding. R. T. Breymeier. *Ibid.*, May 1952.

3. The Determination of Droplet Size in Arc Welding by High Speed Cinematography. P. C. van der Willigen, L. F. Defize. *Philips' Technical Review*, Eindhoven, Holland, vol. 15, no. 4, Oct. 1953.

Cathode Instability in Argon Atmospheres

H. C. LUDWIG
NONMEMBER AIEE

CONSUMABLE steel electrode arcs in argon atmospheres of controlled impurity possess a type of cathode instability which is affected by electrical, chemical, and geometrical parameters. A study of cathode instability and the accompanying fluctuations in arc voltage and current have been made. It is the purpose of this paper to discuss these manifestations as related to the welding of structural carbon steel.

Cathode instability observed as a wandering of the arc flame over the electrically negative terminal of the arc was previously studied by Tanberg[1] for arcs contained in vacua of a few microns of mercury. Tanberg concluded from his experiments of arcs in vacua that the arc potential increased because part of the ionized vapor emitted from the cathode was directed away from the anode or was obstructed from making a direct path to the anode. The same phenomenon has more recently been found by the author to appear with consumable electrode arcs in very pure argon at approximately one atmosphere pressure. In both cases one or another essentially uniform arc potential, intermediate between short- and open-circuit conditions, was found to exist intermittently for varying lengths of time. The prevalency of this condition has been found to be influenced by current magnitude, electrode polarity, purity of arc atmosphere, size of electrode, chemistry of the electrode and of its surfaces, and the chemistry of the workpiece and of its surfaces. Reignition of arcs after electrical short-circuiting also influences the arc potential variation.

The problem of arc instability, using consumable electrodes in argon atmospheres, was also investigated by Doan and his collaborators,[2] who found difficulty maintaining an arc in rigorously purified argon when using purified iron electrodes. However, they found that arcs could be maintained with such 5/32-inch-diameter electrodes when an open-circuit potential of not less than 62 volts was used and a short-circuit current of 110 amperes was exceeded. Arc instability observed as cathode wandering was also reported.

Paper 56-110, recommended by the AIEE Electric Welding Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Winter General Meeting, New York, N. Y., January 30-February 3, 1956. Manuscript submitted March 17, 1955; made available for printing November 22, 1955.

H. C. Ludwig is with the Westinghouse Electric Corporation Research Laboratories, Pittsburgh, Pa.

Cathode Wandering

The cathode wandering observed in the study reported herein is defined as an extension of the arc cathode flame above the tip of the consumed electrode when using straight polarity (electrode negative) or a spreading of the arc cathode flame on the workpiece when using reverse polarity (workpiece negative).

Magnetic oscillographic records taken during these periods of arc irregularity showed that two fairly uniform arc operating voltage conditions existed, Fig. 1(c). These were intermediate in magnitude between arc short-circuit and open-circuit conditions. The per cent of total time of each of these two levels, with respect to total arc time, varied as arc current was changed in magnitude. The per-cent of total time at the high-level arc voltage gradually decreased with increased current when straight polarity was used, Fig. 2, but abruptly changed in magnitude when reverse polarity was used above a critical current of 300 to 325 amperes, Fig. 3.

The wandering of the cathode portion of the arc flame is illustrated by the areas superficially melted on electrode and workpiece when using straight and reverse polarity, Fig. 4. The arcs were contained in a chamber designed to control and maintain arc atmospheres of preselected composition. The electrode, driven automatically as it was arc-consumed, was previously cleaned to remove surface impurities. The workpieces were also surface precleaned. It is appreciated that the more than 99.8-per-cent (%) pure argon gas contains a small percentage of impurities, mainly nitrogen. Also, while the workpieces and electrode surfaces were cleaned, some impurities were still present in minute quantities. The cathode-traversed areas on the electrode and workpiece, which represent the action of the arc when using straight and reverse polarity, were compared with the anode areas for both polarity conditions, Fig. 4. The total wandering of the cathode portion of the arc is obviously large as compared to the anode portion. Cathode wandering was more extensive when using a surface-cleaned electrode and workpiece as compared to uncleaned-surface conditions. Cathode wandering, as evidenced by the superficially melted

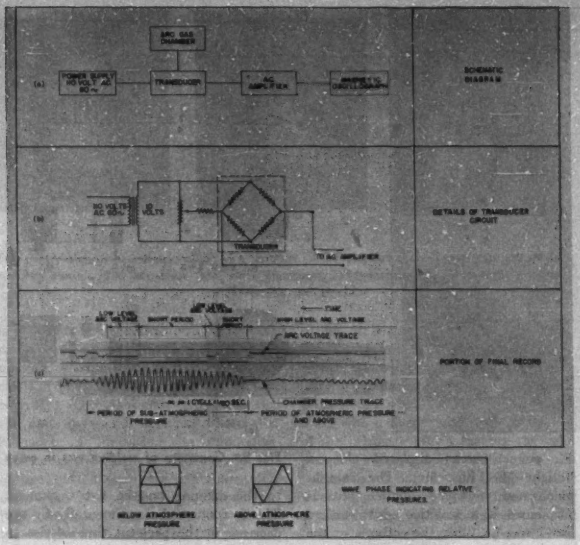


Fig. 1. Pressure-measuring instrumentation and record

areas, increased with arc current when the cathode was on the workpiece. Cathode wandering decreases with increased current when the cathode is on the electrode, although the evidence is destroyed by its consumption.

The wandering of the cathode may be accompanied by fluctuations in arc voltage and current depending on the volt-ampere relationship of the welding power supply used, Fig. 1(c). A constant potential generator with current-limiting resistors was used for these tests at an open-circuit voltage of 70. Tanberg[1] concluded in his experiments of arcs in vacua that a voltage rise occurs because part of the ionized vapor, emitted from the cathode spots, is directed away from the anode or is obstructed from a direct path to the arc anode. The conclusions drawn by Tanberg concerning the arc voltage fluctuation appear applicable to the subject consumable electrode arc. Obstruction of the ionized vapor is possible by the transfer of metal from the electrode to the work, particularly under unstable cathode conditions. Arc-voltage measurements reported herein include current-resistance voltage drops between the available points of measurement and the arc.

Geometrical Considerations

From an analysis of the ionized vapor obstruction condition postulated by Tanberg, one would expect the straight polarity (electrode negative) arc to be more unstable than the reverse polarity arc by consideration of the geometrical parameters of the welding arc, Fig. 5. In the inert gas arc, the anode portion is relatively stable in its position and shall be so considered in this discussion. With reverse polarity, where the area traversed by the cathode is essentially uniplanar, with the electrode maintained in a fixed position and with the anode at its tip, considerable cathode wandering could occur without the arc path being obstructed with respect to the anode. This condition is illustrated in Fig. 5(a) by changes in the location of the cathode from C to C' or to C''. However, a drop of transferred metal could obstruct the arc path with this cathode situation. With straight polarity the cathode would have to move only slightly up the electrode from the C position to be out of direct line with the anode beneath the tip of the electrode on the flat workpiece, Fig. 5(b). Also, the influence of transferred electrode metal would be contributory to ionized vapor obstruction as was the case for reverse polarity.

A second geometrical aspect is illustrated when considering the surfaces over which the cathode travels. Doan,[4] and later the author, considered this to

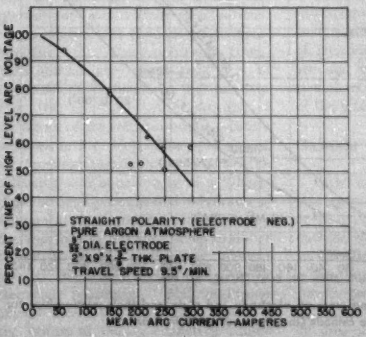

PERCENT TIME OF HIGH LEVEL ARC VOLTAGE
STRAIGHT POLARITY (ELECTRODE NEG.)
PURE ARGON ATMOSPHERE
1/8 DIA. ELECTRODE
2" X 9" X 1/4" THK. PLATE
TRAVEL SPEED 9.5"/MIN.
MEAN ARC CURRENT—AMPERES

Fig. 2 (left). Relation between arc current per-cent time at high-level arc voltage for straight polarity

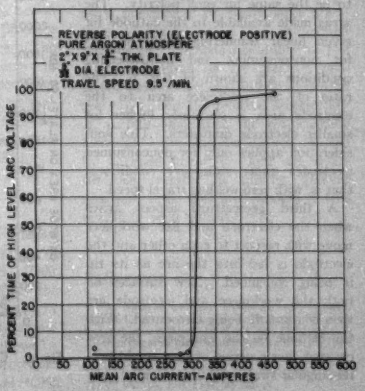

REVERSE POLARITY (ELECTRODE POSITIVE)
PURE ARGON ATMOSPHERE
2" X 9" X 1/4" THK. PLATE
1/8 DIA. ELECTRODE
TRAVEL SPEED 9.5"/MIN.
PERCENT TIME OF HIGH LEVEL ARC VOLTAGE
MEAN ARC CURRENT—AMPERES

Fig. 3 (right). Relation between arc current and per-cent time at high-level arc voltage for reverse polarity

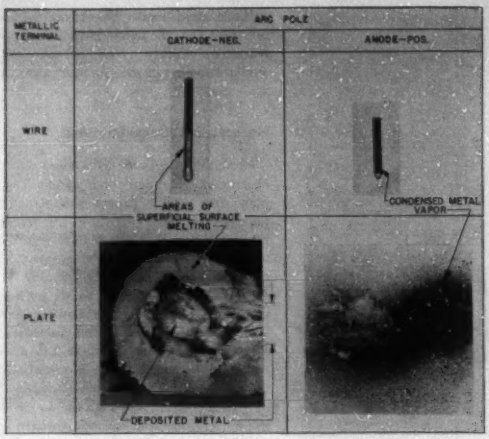


Fig. 4. Cathode and anode areas for pure argon arcs

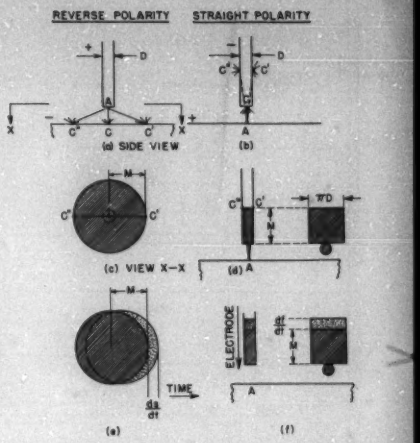


Fig. 5. Geometry of welding arcs in argon

be important with regard to the influence of impurities on the arc poles. It is assumed that the areas on the workpiece over which the cathode flame can freely move are essentially flat surfaces and those on the electrode are cylindrical. The total circular area covered by the cathode when at the workpiece is πM^2 where M is a limit of wandering from the center beneath the electrode, Fig. 5(c). With the cathode at the electrode, the area covered is the total wall surface of a cylinder plus one end and is represented by $(\pi DM)+\pi 4/D^2$ or $\pi D(M+0.25D)$ where D is the diameter of the electrode and M is the linear extent of cathode wandering up the electrode, Fig. 5(d). The M limit is assumed to be the same for each polarity. The areas made available to the cathode for given linear wandering limits of (M) for nonconsuming and nontraversing conditions are shown in Fig. 6. The ratio of the workpiece area to the electrode area is obviously higher for smaller electrode diameters. This consideration applies only for nonconsumed electrodes under stationary conditions, that is, with zero welding travel speed.

A third geometrical aspect arises when the electrode and flat workpiece move with respect to each other and the electrode is fed into the arc as its tip is being consumed. New surfaces on both the workpiece and electrode are then continually being introduced. With the cathode on the workpiece, the area covered over a period of time would be a rectangular swath with semicircular ends, i.e., circular areas displaced along a straight line, Fig. 5(e). The rate at which new workpiece surface area is introduced is a function of linear arc travel speed. With the cathode at the electrode (straight polarity), new end-electrode surface area as well as new cylindrical surface area is continuously being made available, Fig. 5(f).

The consumption rate of a given size electrode increases with current. The rate of consumption can conceivably be sufficiently rapid or the welding travel speed can be sufficiently slow so that the rate at which new surface is made available to the cathode on the electrode may be equal to or greater than the rate at which new surface is made available to the cathode on the flat workpiece. This is confirmed experimentally by the decrease in the per-cent time of rise to a higher arc voltage with higher currents, Fig. 2. With reverse polarity, an increase of arc current and a constant welding travel speed produces an opposite condition wherein the per-cent time for which the higher voltage is experienced increases greatly at about 300 to 325 amperes, indicating insufficient surface impurities are available on the workpiece surface to maintain an arc at the lower voltages and the higher arc currents, Fig. 3. The relation between cathode wandering and fluctuating arc voltages

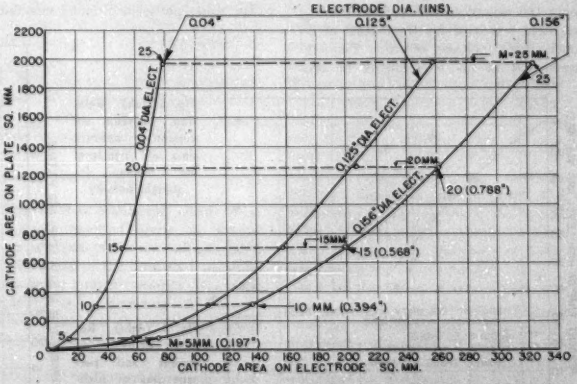


Fig. 6. Calculated cathode-affected areas on electrode and plate and their relationships for movements of the cathode (instantaneous conditions only)

3.5% O_2 1.9% O_2 0.9% O_2

Fig. 7. Variation of cathode climb as a function of the oxygen content of the arc gas shield: welding current, 380 amperes; bare steel electrode, 1/16-inch diameter; straight polarity; argon-oxygen gas shield

is not completely understood; however, certain evidences indicate that the amount of impurity required to decrease cathode instability is a function of arc current. This is particularly true in the case of reverse polarity, where the influence of limitation on impurities present at the cathode with a variation of arc current can be experimentally observed. The transition shown in Fig. 3 indicates that the lower voltage arc cannot be supported above the critical current of about 300 amperes. This critical current may shift to higher values if more impurity is made available to the cathode.

Pressure Variations

One of the more significant effects of rapid fluctuation of voltage and current which accompany cathode instability is to cause gas turbulence in the arc environment which tends to break up the intended argon gas shield and introduce harmful air atmosphere constituents to the liquid weld metal. Nitrogen has been found to produce gross porosity when welding steel. The presence of gas turbulence, caused by arc instability such as electrical short-circuiting and intermittent arc voltage fluctuations, causes intermittent changes in gas pressure, Fig. 1(c). The gas pressure transducer and auxiliary circuit used in making these measurements is illustrated in Figs. 1(a) and 1(b). A photograph of the record of arc voltage and gas pressure is shown in Fig. 1(c). The operation of the pressure transducer is based on the balancing and unbalancing of a Wheatstone bridge by strain and relaxation of component filament pairs which make up the legs of the bridge. The strain or relaxation of the component filament pairs of the bridge circuit changes their electrical resistance. This resistance change alters the electrical balance of the bridge so that a current is caused to flow in the output circuit. The 60-cycle current input used produces a carrier wave form whose amplitude is directly proportional to the applied pressure on a diaphragm of the transducer which is coupled to the filaments of the bridge. Before making the gas-pressure records upon which the arc voltage trace occurs, the bridge circuit was balanced by a potentiometer resistance, Fig. 1(b), to a minimum amplitude of the wave with the prevailing atmospheric pressure condition as observed on the viewing screen of the magnetic oscillograph. Prior to each test, the phase reversal relation of the output wave was determined with pressure above and below the atmospheric pressure. The legend of Fig. 1 shows this relation. The time scale is represented by the vertical lines on the oscillogram record marking time periods of 1/60 second. Immediately following the occurrence of the electrical short circuit, indicated by the sudden drop in voltage to approaching zero, Fig. 1(c), a phase reversal in wave form on the gas-pressure trace occurs indicating a gas pressure below atmospheric. This is a period when surrounding gas at a higher pressure would be drawn into the confines of the subsequently reignited arc. Changes in gas pressure occurring as a consequence of intermittent changes in arc voltage are also in evidence. These are two conditions which cause the contamination of the intended inert gas arc shield.

Gaseous Oxygen Additions

Early efforts were made to produce stable straight polarity arcs by the use of argon-oxygen gas shield mixtures. That cathode wandering was decreased by the introduction of oxygen is shown by Fig. 7 where the amount of oxygen in the argon-oxygen gas shield was varied, as shown by the three composite sketch-photographs of arcs using 3.5%, 1.9%, and 0.9% oxygen in the argon-oxygen gas shield. In the first case, the arc extends only slightly above the tip of the electrode. As the oxygen content of the gas shield is decreased, cathode wandering up the electrode is increased. Under all three conditions of operation illustrated in Fig. 7, the arc cathode was stable in position, but at different locations. With the use of 3.5% oxygen and 1/16-inch-diameter electrode at 380 amperes, the cathode was sufficiently stable so that voltage fluctuations previously described were absent. These variations were present with the lesser amounts of oxygen shown. That this is true, except in instances of arc reignition after arc interruption caused by electrical short-circuiting, is evidenced by ultraslow motion pictures. Ultraslow motion pictures are useful to show the influence of the argon-oxygen gas shield upon the welding arc.

Summary

It has been shown that the consumable steel electrode arc in argon atmospheres of controlled impurity may exhibit a type of instability that is manifested by: (1) wandering of the cathode, and (2) rapid arc voltage fluctuations. The manner in which arc voltages vary with current have been evaluated for each polarity. The per cent of time that the voltage rises to the higher values decreases with increasing current for straight polarity. With reverse polarity, the arc voltage remains at a nearly constant value until some apparently critical value of current is used, whereupon the voltage rises to a higher operating value. It is unlikely that this condition will occur in practical welding because of the usual presence of relatively large quantities of impurities on the production workpieces. Cathode instability of the arc is influenced by arc polarity, geometric parameters, obstruction of ionized arc vapor, and the relative accessibility of surface impurities on the cathode terminal. Arc gas turbulence, caused by cathode instability and arc interruptions, results in mixing gaseous nitrogen into the arc stream. This results in weld porosity.

References

1. A Theory of the Electric Arc Drawn Under Low Gas Pressure, R. Tanberg. Westinghouse Research Scientific Paper 405 Westinghouse Electric Corporation, Pittsburgh, Pa., Dec. 26, 1929

2. Arc Welding in Argon Gas, G. E. Doan, W. C. Schulte. *The Welding Journal*, New York, N. Y., vol. 15, Jan. 1936, pp. 23–27.

An Automatic Output Regulator for Power Oscillators

E. G. HOPKINS
NONMEMBER AIEE

H. J. FRASER
NONMEMBER AIEE

Synopsis: The paper describes a new system of automatic output regulation for induction heating oscillators. With the use of a power triode as part of the oscillator grid-resistor an oscillator is made to block whenever a negative voltage is applied to the triode grid. This grid voltage is applied for a controlled portion of each half-cycle by means of two small thyratrons whose firing angle is determined by the phase of an error signal. The error signal is derived from a tungsten lamp bridge in which the lamp is heated by part of the high-frequency output of the oscillator. This system enables an existing oscillator to be regulated without alteration of its power supply. An automatic regulator unit using the system is described. It is interchangeable between a number of oscillators of different designs installed in the exhaust section of a tube plant.

THE evacuation of receiving tubes is carried out on automatic machines having a number of exhaust ports distributed around a circular horizontal turret which rotates with an intermittent motion. The electrodes of the tubes are induction heated[1] by coils which are usually connected in series to form part of the tank circuit of an oscillator of several kilowatts' rated output. The current in the coils must be maintained close to some experimentally determined optimum value and in particularly critical exhaust schedules even a few per cent variation can have a marked effect on tube quality.

The output current of the oscillator is influenced by variations in supply voltage and by variations in the number of tubes situated within the heating coils. At the commencement and end of a production run when the machine is partially filled with tubes, large changes in current may take place if the oscillator is not adjusted to compensate for the changing load. Significant changes may also occur whenever the steady flow of tubes from the sealing machine is interrupted by the presence of faulty tubes. During the indexing period, the coils are raised to allow the tubes to pass below them, leaving the oscillator completely unloaded. The response time of any system of oscillator regulation must therefore be short compared with the period of dwell if effective control is to take place over the greater part of the heating time. In high-speed exhaust schedules the period of dwell may be as low as 5 seconds.

The conversion of an already existing oscillator to automatic output regulation with short response time would normally involve the rebuilding of the power supply to incorporate thyratron control. The exhausting section of a tube plant is likely to contain a variety of oscillator designs, each of which would have to be treated as an individual problem when converting it to automatic regulation. The present paper describes an automatic output regulator incorporating a novel principle of output control which enables any oscillator to be regulated without alteration to its existing power supply.

In the new system of control two parallel power triodes in series with the oscillator grid resistor act as switches to isolate this resistor from the filament of the oscillator tube during a controlled portion of each half-cycle of mains voltage. The oscillator thus can be made to block for a variable fraction of the total time. This system achieves the same result as a previous system of manual control developed by Babat and Losinsky,[2] but has the advantage that the negative bias required to prevent oscillation is supplied, not by a separate rectifier but by the oscillator itself. The voltages applied to the grid of the switching triodes can be derived from a receiver-type rectifier and two small thyratrons controlled by a low-power error signal. The cost and weight of a regulator using the system can therefore be kept comparatively low.

By the use of a small number of regulator units, which are interchangeable between one oscillator and another, it is possible to provide automatic output regulation whenever this is required at a small fraction of the cost of converting all the oscillators in a tube plant to automatic regulation.

Regulator System

The automatic regulator system is represented by the block diagram of Fig. 1. It consists of the measuring circuit, the controller, and the oscillator. The measuring circuit converts changes in rms bus bar current to changes in the phase of a mains frequency error signal. The controller and oscillator convert these phase changes to changes in rms bus bar current. Such a closed system will compensate for variations either in mains voltage or in load.

The controller circuit is shown in Fig. 2. The rectifier V_1 supplies a negative voltage developed across R_1 of the divider R_1-R_2 to the grids of the switching triodes V_2-V_3 through the resistor R_3. R_3 is also in the cathode circuit of the thyratrons V_4 and V_5. At some point in each half-cycle one thyratron fires, cancels the negative voltage at the grids of the switching triodes, and permits the oscillator tube to oscillate for the remainder of the half-cycle. At the end of the half-cycle the thyratron ceases conduction and the switching tube is again cut off, causing the oscillator to block. The firing point of the thyratrons is controlled by the phase of the error signal.

The measuring circuit, shown in Fig. 3, consists of a tungsten lamp bridge which produces a phase-modulated mains-frequency error signal and a tuned limiting amplifier which converts the signal into a constant amplitude sine wave of 70 volts rms. The bridge has three fixed equal resistive arms, the fourth being a tungsten lamp connected to the bridge through an inductance-capacitance low-pass filter. The lamp is fed with a high-frequency current by means of a variable coupling air-core transformer, T_1, whose primary winding is connected between two tappings on the high frequency bus bar. A capacitor C_1 isolates the transformer from the 50-cycle-per-second bridge circuit. Because the reactances of the bus bar and the transformer are much greater than their own resistances or the resistance due to the lamp, the high-frequency current in the lamp is proportional to bus bar current irrespective of fluctuations in oscillator frequency. The mains-frequency current flowing in the lamp is 20 per cent (%) of the high-frequency current and contributes 2% of the total rms current. Consequently, the bridge may be assumed to balance

Paper 56-15, recommended by the AIEE Electric Heating Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Winter General Meeting, New York, N. Y., January 30–February 3, 1956. Manuscript submitted August 30, 1954; made available for printing October 31, 1955.

E. G. Hopkins is with the New South Wales University of Technology, Sydney, Australia, and H. J. Fraser is with the Amalgamated Wireless Valve Company, Ltd., Sydney, Australia.

The authors are indebted to the Amalgamated Wireless Valve Company, Pty. Ltd., of Sydney, Australia, for permission to publish this paper.

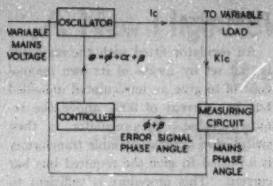


Fig. 1. Block diagram of automatic regulator system

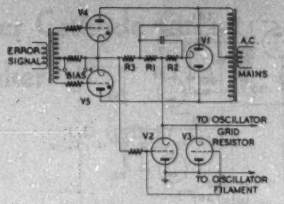


Fig. 2. Controller circuit

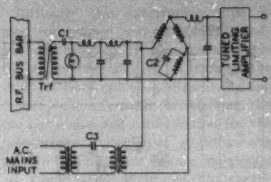


Fig. 3. Measuring circuit

at some particular value of bus bar current for each setting of the variable transformer, independently of fluctuations in mains voltage. The capacitor C_3 adds a small fixed quadrature voltage to the bridge output causing it to vary in phase as well as amplitude.

Characteristics of System

The characteristics of both the measuring circuit and the controlled oscillator are nonlinear functions. The angle-current characteristic of the measuring circuit may be expressed as

$$\phi = \tan^{-1} \frac{0.63(I_c - I_b)}{R\omega C_3 I_b} + \frac{\pi}{2}$$

where ϕ is the phase angle between the output and the input of the bridge R is the resistance of one arm of the bridge, I_c is the controlled bus bar current, I_b is the bus bar current at which the bridge balances, and ω is the angular frequency of the mains. The gain of the measuring circuit $d\phi/dI_c$ may be increased at any particular value of $(I_c - I_b)/I_b$ by a decrease in the capacitance of C_3. However, because the minimum output voltage of the bridge decreases proportionally with C_3 and must have sufficient amplitude to cause limiting in the amplifier, a limit is placed on the maximum possible measuring-circuit gain.

The foregoing equation may be obtained by analysis of the bridge after making the following assumptions:

1. Lamp current is proportional to I_c.
2. $1/\omega C_3 >> R$.
3. The lamp obeys the law[3] $V = I^{1.43}$.
4. The bridge is unloaded.

Single-phase oscillators, whether self-rectifying or having separate rectifiers, usually produce an output current fully modulated at twice the mains-frequency while 3-phase oscillators have very little output modulation. Assuming sinusoidal initial modulation, the current phase-angle characteristic of single-phase oscillators can be expressed approximately:

$$I_c = I_u \sqrt{1 - \frac{\theta}{\pi} + \frac{\sin 2\theta}{2\pi}}$$

where θ is the phase angle between the commencement of a half-cycle of mains voltage and the commencement of oscillation, and I_u is the current which would flow with the same conditions of oscillator adjustment and no interruption of the oscillator grid circuit. Three phase oscillators obey the law

$$I_c = I_u \sqrt{1 - \frac{\theta}{\pi}}$$

assuming that no initial modulation takes place. The gain of a controlled oscillator $(dI_c/d\theta)$ at any particular value of θ and I_u is not subject to alteration by design. Fig. 4 shows the characteristics of the measuring circuit and the controlled oscillator drawn with their phase scales adjacent and displaced from one another by a constant angle $(\alpha+\beta)$. The angle α is the phase difference between the commencement of a half-cycle of error signal and the commencement of oscillation, and is a function of thyratron bias and error signal amplitude. The angle β is the phase difference between the mains voltage and the input to the bridge. It is set by the capacitance of C_3 to allow the most sensitive section of the measuring circuit characteristic to coincide with the most useful section of controlled oscillator characteristic. The point of correspondence between the current scales in the two diagrams is dependent on the setting of I_u relative to I_b. The effective steady-state loop gain of the experimental regulator system over a typical operating range is −28. Variations in output current may therefore be reduced to about 3% of their uncontrolled values.

The time delays in the system are due to: 1. the thermal inertia of the lamp, 2. the response of the tuned amplifier to rapid changes in phase, and 3. the finite time taken for the

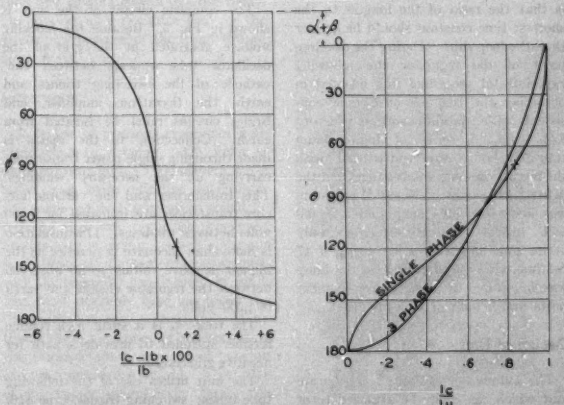


Fig. 4. Steady-state characteristics of the measuring circuit and of single-phase and 3-phase controlled oscillators

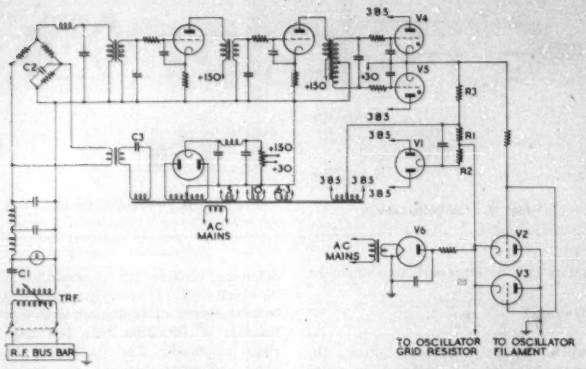


Fig. 5 (above). Circuit diagram of automatic regulator

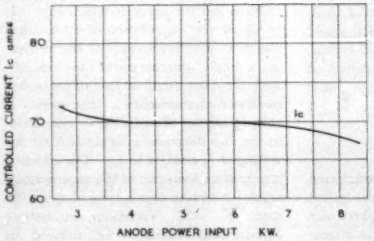


Fig. 6 (left). Regulated output current as a function of anode circuit input power for varying oscillator load

thyratrons to respond to changes in the phase of the error signal. Of these, the thermal inertia of the lamp lends itself to the widest design variations. A simple criterion for stability[4] in a closed loop system having three time constants is that the ratio of the longest to the shortest time constant should be greater than the loop gain. During the development of the regulator the following experimental procedure was adopted in obtaining the best response time consistent with stability and satisfactory loop gain. A series of high-vacuum tungsten lamps was constructed with the use of wires of various gauges and the lightest gauge which has stable operation was adopted. This lamp had a 0.008-inch helical filament supported only at its ends and with a lit length of 17 centimeters. Thermal drift in the lamp was minimized by using heavy filament posts and a bulb of large diameter.

Design of Unit

The automatic regulator is a separate unit which can easily be attached to or removed from any oscillator. Each oscillator is fitted with a large shockproof socket which inserts the switching tubes in place of a 5,000-ohm section of the grid resistor when the regulator plug is in place. The high-frequency connection is made by two clamps attached to the bus bar with a spacing of about 3 feet between them.

The complete circuit of the unit is shown in Fig. 5. Because the blocking voltage generated at the grid of the oscillator tube appears between the cathode of the switching triodes and earth, the thyratron, amplifier, and bridge circuits must be isolated from earth. Connection to the mains is made through a single power transformer carrying all the necessary windings. This transformer and the variable air-core transformer are insulated for 5,000 volts between windings. This insulation is more than adequate in practice as the highest negative voltage so far observed between the regulator chassis and earth is 1,500 volts.

The tube V_6 is a small high-voltage rectifier included to provide a path for negative grid current.

The unit makes use of the following tube types; switching triodes, type *810;* thyratrons, type *2050;* amplifiers, type *6J5-GT;* rectifiers, types *5Y3-GT* and *2X2.*

Performance

An oscillator fitted with the regulator is first set by means of its own manual control to give an unregulated unloaded bus bar current of 30% above the required value. The regulator is then switched in and the variable transformer is adjusted to give the required bus bar current. This procedure is sufficient to ensure that the regulator operates near its region of maximum sensitivity. Fig. 6 shows the measured regulated current of a single-phase oscillator as a function of anode input power with varying load. It will be seen that the variation in the controlled current is 1% over a ratio of input powers of two to one. This is a greater change in loading than that which accompanies the filling or emptying of an exhausting machine. The response time of the oscillator to such changes is of the order of 0.3 second.

The regulator is not affected by the strong high-frequency fields and the frequent line surges encountered in a tube plant where a number of oscillators are being used in close proximity. This freedom from interference is thought to be due to a combination of the following factors: 1. the choice of a phase-modulated amplitude-limited error signal, 2. the large thyratron bias and grid signal, 3. the tuning of the amplifiers; 4. the high-frequency by-pass capacitors at the amplifier and thyratron grids, and 5. the shielding of the amplifier and thyratron circuits. The regulator unit has given satisfactory service with a wide variety of oscillators. It was put into service in 1948 and is at present in regular use. During the intervening period its only servicing has been the replacement of the thyratrons.

Conclusion

The system described in this paper offers a satisfactory solution to the problem of output regulation for existing oscillators and is a possible alternative to thyratron-controlled power supplies in the design of new oscillators.

References

1. Induction Heating in Radio Electron-Tube Manufacture, E. E. Spitzer. *Proceedings,* Institute of Radio Engineers, New York. N. Y., vol. 34. Mar. 1946, pp. 110w–115w.

2. Power Oscillators, G. Babat, M. Losinsky. *Wireless Engineer,* London, England. vol. 17, Jan. 1940, pp. 16–18.

3. Theory of Non-Linear Bridges in Voltage Stabiliser Circuits, G. N. Patchett. *Proceedings,* Institution of Electrical Engineers, London, England, vol. 93, pt. III, Jan. 1946, p. 17.

4. Contributions to the Theory of Automatic Controllers and Followers, D. G. Prinz. *Journal of Scientific Instruments,* London, England, vol. 21. Apr. 1946, pp. 53–63.

A New High-Power Cathode Circuit Breaker

L. D. McCONNELL
ASSOCIATE MEMBER AIEE

J. D. FINDLEY
ASSOCIATE MEMBER AIEE

Synopsis: Recent increases in the power and voltage of d-c systems fed by mercury-arc rectifiers on some installations have created a need for a faster and larger capacity cathode circuit breaker (Fig. 1). A new breaker has been developed for this severe service.

TODAY'S highest power d-c systems have many large rectifier units operating in parallel. These systems are subject to a fault condition peculiar to mercury-arc rectifiers and generally known as arc back. This fault is caused by the occasional inability of an anode and cathode pair to withstand the inverse voltage during the normally nonconducting portion of the cycle. Reverse or arc-back current then begins to flow. Once started, it increases rapidly. The ultimate current is limited by the arc drop of the faulted anode-cathode pair, by the resistance of the reverse current path, and by the regulation of the remainder of the d-c system. As the d-c system is large and the resistances in the reverse path are low, the ultimate current is relatively very large.

This ultimate arc-back current is much beyond that which might be tolerated, even momentarily, by the rectifier equipment because of the mechanical stresses and overheating involved. When used for arc-back protection, a cathode breaker must be fast enough to limit the current to a tolerable value. The safe and continuous operation of such concentrated systems depends upon the removal of the d-c short circuit by the breaker, and so the breaker must be reliable. This reliability must include the ability to clear maximum rate-of-rise arc-back currents many times coupled with the minimum need for inspection and maintenance.

Paper 56-185, recommended by the AIEE Industrial Power Rectifiers Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Winter General Meeting, New York, N. Y., January 30–February 3, 1956. Manuscript submitted October 20, 1955; made available for printing December 7, 1955.

L. D. McCONNELL is with the Canadian Westinghouse Company, Ltd., Hamilton, Ont., Canada, and J. D. FINDLEY is with the Westinghouse Electric Corporation, East Pittsburgh, Pa.

There is a trend toward larger rectifier unit ratings, both in voltage and current. Hence the rating of the new breaker being discussed herein was established as follows:

Voltage—1,200 volts.
Current—10,000 amperes continuous.
Interrupting ability—In a system in which arc-back cathode current might rise to 100,000 amperes in 0.008 second if no breaker were in the circuit, the breaker should limit the current to 75,000 amperes. Interruption of such a fault should be completed in one cycle.

To understand better the performance required of the breaker, an analysis of the usual double 3-phase circuit was made from the point of view of the action of the breaker voltage in the circuit. An initial effective resistance for the first half-cycle in the arc-back current path of 0.02 ohm was assumed. Further, it was assumed that current limitation might be required soon after the instant of the first peak inverse voltage. For such a condition, the breaker arc-drop voltage must be approximately 1,000 volts at the limitation instant for the current to be limited to 75,000 amperes. This condition is somewhat more severe than any circuits known to exist today, but was set as a desirable goal for the interrupter because of the trend toward larger units.

Interrupter

A review of previous experience indicated that to develop 1,000 volts or more arc drop on a 75,000-ampere arc, a larger and more effective interrupter and faster contact separation would be required than any previously employed on this type of circuit breaker. Various basic types of interrupters were studied, with the reliability requirement in mind. The most desirable type employed the "magnetic de-ion" principle.[1] Many thousands of a-c power air breakers using this principle are in service and have an excellent experience record. This type of interrupter magnetically forces the arc upward into the inverted "V" slots of a series of spaced zircon porcelain plates.

Although this interrupter type is most widely used on a-c breakers, standard a-c interrupters have considerable d-c interrupting ability. A comparative study of test results on various existing interrupters was made in selecting the arc chute size and configuration for the new breaker. Particular attention was paid to maximum arc voltage, maximum arc power, and integrated arc energy found by test, as compared to the circuit requirements discussed in the foregoing.

To co-ordinate with the arc chute size and configuration selected, an air-core blowout magnet was chosen. The blowout coil is so shaped as to give roughly uniform field distribution over the whole interrupting area of the arc chamber. See Figs. 2 and 3. To insure earliest possible motion of the arc off the contacts, the coil is connected permanently in series with the breaker current. To aid further the rapid motion of the arc as it is drawn on the contacts, a laminated open iron ring is arranged around the blowout coil with a gap at the contacts. This provides increased field intensity in the contact area at the low currents existing when the contacts begin separating. The arc horns above the contacts and the blowout coil are co-ordinated in shape so that good field strength is provided at the arc terminals to move those terminals or hot spots rapidly upward as the arc lengthens and enters the slots. This permits the contact area to withstand the relatively high voltages developed when the arc is high in the arc chute at the end of the interruption.

Fig. 1. Side view of cathode circuit breaker

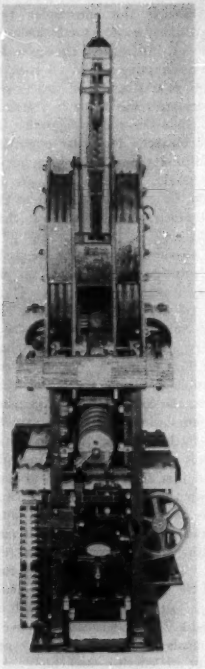


Fig. 2. Front-end view of breaker arc chute tilted back for inspection of contacts

Contacts

In designing the contact system, there were certain primary considerations. Durability is essential to enable the contacts to withstand the severe duty with minimum need for inspection and maintenance. Load-carrying ability must be assured. Earliest possible contact separation is required to start the limitation of fast-rising fault currents. Maximum opening acceleration is needed to obtain early arc lengthening and also to withstand the high voltage across the contacts during and after the interruption. Both early parting (when the current is smaller) and high acceleration help to reduce the erosion of the contacts.

To obtain main contact parting as early as possible, the resilience for the main contacts was put in the stationary contact. This resilience is necessary to assure contact pressure when contacts wear and adjustments change. The ratio of the accelerating force (contact pressure springs) to the mass of the movable portion of the stationary main contacts is made small as compared to the same ratio in the moving contact with the initial motion of the moving contact system. This causes early transfer of the current from the main contact path to the arcing contact path. Since there is no immediate main contact follow, the amount of arcing contact follow needed to assure full protection of the main contacts from arc damage is less by the distance formerly allowed for main contact follow.

Table I. 750-Volt D-C Tests, Four 1,500-Kw Generators in Parallel

Ultimate Current, No Breaker: 80,000 Amperes; Inductance: Approximately 70 Microhenries

	Test Number			
	10	34	35	36*
Max. current, amperes	36,000	35,000	36,000	36,000
Time to limit, second	0.0065	0.0063	0.0066	0.0067
Time to clear, second	0.0093	0.0095	0.0096	0.0095
Max. arc drop, volts	2,200	1,900	2,100	1,900
Max. arc power, megawatts	21	22	32	27
Total arc energy, kw-sec	57	57	72	68

* Oscillogram, Fig. 4.

To reduce moving contact mass, water cooling is used in the moving main contacts. The stationary mains are also water cooled. Solid inserts of conventional silver-nickel alloy are used on the main contacts. To assure durability and moderate erosion, the arcing contacts are generously proportioned. Contact tip material is tungsten silver. Moving and stationary arcing contacts are identical to simplify renewal. Both are readily removable without dismantling any other components. These factors combine to afford long life with minimum cause for readjustment or replacement, yet permit ease of maintenance when required.

Three simple contact adjustments are provided: stationary main contact travel, moving contact closed position, and arcing contact follow. Settings are easily made and readily checked by visual inspection. There is no need for contact pressure measurement or adjustment.

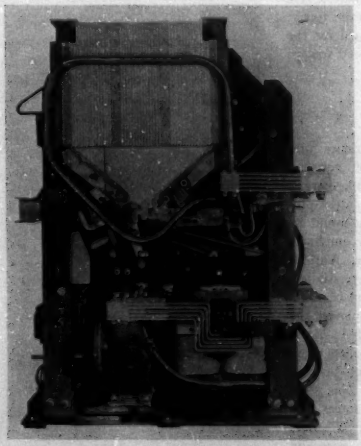


Fig. 3. Arc chute interior shown by removing side plates

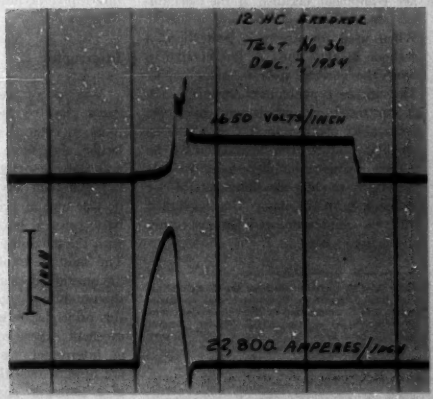


Fig. 4. 750-volt d-c test

Table II. 1,500-Volt D-C Tests, Four 1,500-Kw 750-Volt Generators in Series-Parallel

Ultimate Current, No Breaker: 40,000 Amperes; Inductance: Approximately 280 Microhenries

	Test Number		
	20	31*	32
Max. current, amperes	23,400	23,600	24,600
Time to limit, second	0.0064	0.0073	0.0074
Time to clear, second	0.0109	0.0115	0.0111
Max. arc drop, volts	4,200	4,000	4,100
Max. arc power, megawatts	35	27	24
Total arc energy, kw-sec.	86	98	75

* Oscillogram, Fig. 5.

Holding Magnet

As is usual with high-speed circuit breakers of this type, the contacts are held in the closed position by a holding magnet and opened by a spring. However, several new departures are made in the holding magnet design. The opening acceleration obtainable in a system retained in the closed position by a holding magnet, is determined by the ratio of the holding force on the armature to the inertia of the moving system. The mass of the armature is a considerable part of the system inertia. Hence maximum ratio of armature holding force to armature mass is desired. The maximum force per square inch of armature face obtainable with silicon steel is about 200 pounds. This can be obtained only with very high magnetizing forces. However, a special iron cobalt alloy[3] develops a flux density over 23,000 gauss with a moderate magnetizing force of 200 oersteds. This results in a holding force of about 300 pounds per square inch and makes possible a practical armature design weighing less than one pound but retaining a spring force over 900 pounds. This was an important factor in obtaining an initial arcing contact acceleration of over 150 g.

Another departure from past practice in the holding magnet design is the method of releasing the armature. Experience indicates that most arc backs occur at the beginning of the nonconducting period, and cause reversal of cathode current at rates from 5 to 10 amperes per microsecond. However, a small percentage may occur at a later portion of this period when voltage to cause reversal is larger. Under such a condition extremely rapid reversal might occur on the larger power systems. To prevent any possibility of developing retarding forces on the tripped armature with such extremely rapid reversals, the main current is not passed through the holding magnet as in past practice. Instead, a tripping or diverting magnet is employed. The cross section of this diverting magnet is so proportioned that when saturated it will carry only the flux originally passing through the armature. The effect of the main current flux is thus limited to that required positively to release the holding armature.

Interrupting Performance Tests

To prove interrupting ability in the laboratory, four types of tests were made. First, the circuit breaker was checked by opening short circuits on four 750-volt 1,500-kv d-c generators in parallel. The ultimate short-circuit current on these machines is 80,000 amperes.

Table III. 12-Cycle Tests, Direct Connection to Generator

Approximate Circuit Constants: Crest Current, No Breaker, 50,000 Amperes; Crest Voltage, 1,650; Average Rate of Rise, 4.2 Amperes per Microsecond; R, 0.017 Ohm; L, 370 Microhenries

	Test Number		
	FK	FL	FM*
Max. current, amperes	34,300	34,500	32,000
Time to limit, second	0.0082	0.0086	0.0076
Time to clear, second	0.0117	0.0124	0.0115
Restored volts, volts	1,650	1,650	1,650
Max. arc drop, volts	5,300	5,200	5,200
Max. arc power, megawatts	93	92	84
Total arc energy, kw-sec.	280	270	230

* Oscillogram, Fig. 6.

The breaker easily limited this to about 35,000 amperes in 0.0065 second. About 60 tests of this type were made. A few representative tests are shown in Table I and a typical oscillogram is shown in Fig. 4.

Next, the breaker was tried on 1,500 volts with the same four generators connected series-parallel. In this case the ultimate short-circuit current is 40,000 amperes. The breaker limited this to about 24,000 amperes at 0.0075 second. About 25 tests of this type were made. Three typical tests are shown in Table II and an oscillogram in Fig. 5.

To obtain laboratory test power beyond the limits of any available d-c facilities, a-c high-power equipment was used in two ways. First, direct connection was made to a large alternator turning at low speed to produce approximately a 12-cycle voltage. To simulate

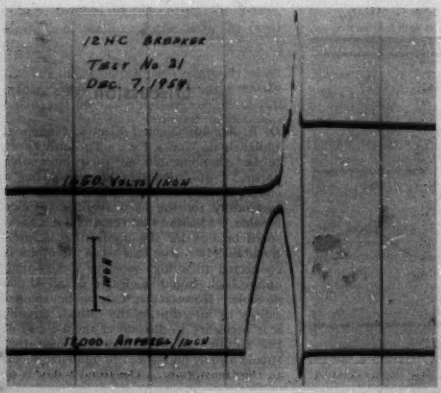


Fig. 5. 1,500-volt d-c test

Table IV. 25-Cycle Tests Using Three High-Current Transformers in Parallel

Approximate Circuit Constants: Crest Current, No Breaker, 108,000 Amperes; Crest Voltage, 1,420; Average Rate of Rise, 12.4 Amperes per Microsecond; R, 0.0045 Ohm; L, 80 Microhenries

	Test Number		
	IY*	IZ	JA
Max. current, amperes	76,000	77,000	77,000
Time to limit, second	0.0068	0.0077	0.0077
Time to clear, second	0.0110	0.0118	0.0127
Restored volts, volts	675	600	500
Max. arc drop, volts	2,000	2,800	2,800
Max. arc power, megawatts	131	142	107
Total arc energy, kw-sec.	450	500	530

* Oscillogram, Fig. 7.

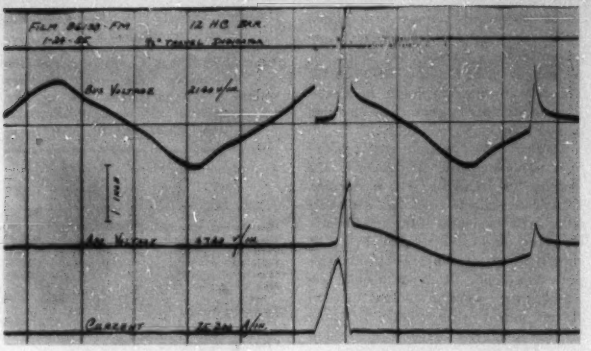


Fig. 6. 12-cycle test direct connection to generator

direct current with this a-c laboratory testing, a synchronous timing device was used to cause the short-circuit closing switch to close at a chosen instant on the a-c wave. This instant was in the half-wave with correct polarity to cause reverse current in the breaker. The breaker requires in general about 12 milliseconds to interrupt. The choice of location of these 12 milliseconds on the 12-cycle half-wave is necessarily a compromise between initial voltage and restored voltage.

About 140 tests were made with this 12-cycle method. For the three tests shown in Table III, timing was adjusted to favor high restored voltage at some sacrifice in rate of rise. Note in Fig. 6 that the breaker easily develops sufficient arc voltage quickly to complete the interruption in this rather inductive circuit. This indicates that the interrupter is able to withstand ample restored voltage.

In a second method of testing with high-power a-c equipment, one phase of a large alternator turning at 25-cycle speed was connected to the high-voltage windings of three special transformers in parallel. The low-voltage windings of these transformers are designed for very high currents. The necessary closing switch and backup breaker were connected in the high-voltage leads. Connections from the transformers to the test breaker could then be made short and heavy to give an effective resistance in the test breaker circuit of about 0.0045 ohm. This unusually low resistance forced the test breaker to develop high arc voltage at current limitation. Again, a synchronous timing device was used to iniate the short circuit. This time, however, the timing was adjusted to give high rates of rise and the test breaker was forced to develop high arc voltage at high currents.

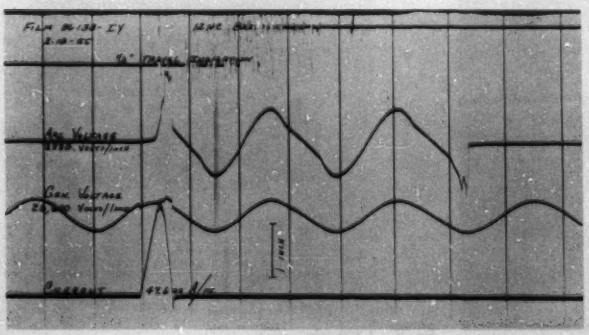


Fig. 7. 25-cycle test using high-current transformers

About 90 tests of this 25-cycle type were made, three of the larger tests are shown in Table IV. These tests proved both the breaker's speed of action and the ability of the interrupter to handle the high power and energy resulting from high arc voltage at high current. Note the magnitude of the quantities in Table IV. Note also in Fig. 7 that at the instant of limitation the arc voltage is nearly 1,000 volts while a 76,000-ampere arc exists. This unusually large power test of a d-c interrupter combined with the previous tests at higher voltage on a more inductive circuit give assurance that the interrupter will be adequate for the stated rating.

In all these laboratory tests, only one set of main contacts was used. Wear was negligible. Over 150 high power tests were made on one pair of arcing contacts. The useful life of this pair of arcing contacts was only partially exhausted by these tests. During the testing, slight readjustments of the contacts were made only at intervals of many tests.

Conclusion

The new cathode breaker has been shown by test to have the following features which are very necessary for this high-power service:

1. Early current limiting.
2. Reliable interrupting performance in repetitive duty.
3. Durability to minimize maintenance.

References

1. Magnetic "De-ion" Air Breaker for 2,500–5,000 Volts, L. R. Ludwig, R. H. Nau. *AIEE Transactions*, Sept. 1940, vol. 59, pp. 518–22.
2. Hiperco—A Magnetic Alloy, J. K. Stanley T. D. Yensen. *Ibid.*, vol. 66, 1947, pp. 714–18.

Discussion

O. B. Arnold (General Electric Company, Philadelphia Works, Pa.): The authors are to be complimented on their interesting paper. I am somewhat surprised that a circuit breaker of this type was developed specifically for use as a rectifier cathode breaker. I believe that the need for a high-speed breaker for this application is questionable. It is true that when a rectifier is connected to a high-capacity d-c system, an arc back should be cleared as quickly as possible. However, it cannot be cleared simply by interrupting the cathode circuit; it is necessary to disconnect the a-c end of the rectifier as well, and this operation should also be high-speed to minimize shock to the transformer. The trend today is to

use high-speed anode breakers for this purpose, in conjunction with semihigh speed cathode breakers. The anode breakers alone can clear the arc back without the necessity of opening the cathode circuit, and they have the advantage of disconnecting only the tank in which the arc back occurred, permitting the remainder of the rectifier unit to continue operation. The addition of high speed to the cathode breaker therefore adds nothing to the rectifier performance or protection, and results in the loss of service continuity that the anode breaker would otherwise provide. Anode breakers are available for use on d-c systems with voltages up to 1,000. Whenever higher direct voltages are required, we believe that the proper approach to the problem would be the development of anode breakers of higher capacity, rather than the use of high-speed cathode breakers.

Regarding the mechanical design of the circuit breaker, I wonder if the increase in speed obtained by the use of water-cooled contacts is worth the added complications. The water-cooling system with its circulating pump, heat exchanger, and control equipment is a source of possible trouble and increased maintenance. Also, the difference in potential between various sections of the cooling system can result in damage due to electrolysis.

L. D. McConnell and J. D. Findley: In answer to Mr. Arnold, we would emphasize that the magnitude of the current at the instant of arcing contact-parting largely determines the rate of erosion of the contacts, and the consequent need for maintenance. Also the size of the current at the time the arc moves up into the interrupter greatly affects the necessary size of the interrupter. To achieve the desired limitation the arc must be off the contacts and up into the interrupter early enough to get the necessary arc drop voltage. Thus contact speed is valuable in several ways.

In many rectifier installations the water-cooling system can include the circuit breaker without much added complication. Electrolysis and condensation problems can be adequately handled by careful design.

On the other hand, preliminary studies indicated that an air-cooled contact design would be much larger, more costly, and probably about 0.002 second slower at contact parting. This would mean increased erosion of contacts and interrupter and would result in current-limiting outside of the design figure.

Since the final proof of a new design is in actual service experience, the authors are thankful for the comments about good service, particularly as to maintenance.

Methods of Measuring Arc-Drop Voltage on Mercury-Arc Rectifiers

HAROLD WINOGRAD
FELLOW AIEE

WILLIAM E. LAWTON
MEMBER AIEE

THE over-all efficiency of a mercury-arc rectifier unit is determined by the segregated loss method. The losses in the component parts of the unit—the rectifier transformer, rectifier, and its auxiliaries—are measured separately, and the efficiency is calculated. The methods used for measuring the losses are specified in the American Standards Association (ASA) Standard for Pool Cathode Mercury-Arc Power Converters.[1] This method is used for determining the efficiency, instead of measuring the power input and output of the unit, for two reasons:

1. The losses constitute a relatively small percentage of the unit rating (approximately 5 per cent for a d-c output voltage of 600 volts) and can be determined more accurately by individual loss measurements.

2. To set up a complete rectifier unit in the factory for a full-load test would be costly and is not always practical, because of the limitations of available testing facilities.

Paper 56-187, recommended by the AIEE Industrial Power Rectifiers Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Winter General Meeting, New York, N. Y., January 30-February 3, 1956. Manuscript submitted October 3, 1955; made available for printing December 1, 1955.

HAROLD WINOGRAD is with the Allis-Chalmers Manufacturing Company, Milwaukee, Wis., and WILLIAM E. LAWTON is with the Aluminum Company of America, Vancouver, Wash.

The arc-drop loss in the rectifier tubes is usually the major part of the total unit losses. It constitutes about 60 to 80 per cent of the total losses for d-c output voltages in the range of 750 to 250 volts. For this reason considerable attention has been given to the methods used for measuring the arc-drop voltage.

Over the years, various methods have been employed for measuring the arc-drop loss.[2] These may be classified in general under the following categories:

1. *Heat Dissipation Method.* The power loss in the rectifier is determined by measuring the heat absorbed by the rectifier cooling water, and estimating the additional heat dissipated from the rectifier surfaces by convection and radiation.

2. *Wattmeter Input-Output Method.* With the rectifier operating at a low voltage (below 100 volts), the power input and output are measured, and the rectifier loss is determined by subtracting the power output and transformer loss from the power input.

3. *Direct Wattmeter Method.* The power loss in a rectifier tube is measured directly by means of a wattmeter, with its current coil connected to the anode current circuit, and its voltage coil in a circuit between the anode and cathode terminals.

4. *Oscillograph Method.* An oscillogram is made of the arc-drop voltage, and its average value during the anode conducting period is determined by measurements on the oscillogram.

In recent years, methods 3 and 4 have become the generally accepted test methods. In the United States the oscillograph method has been used and is specified in the ASA *C-34* standards. In Europe both methods have been employed. In the Swedish standards for mercury-arc rectifiers, issued in 1941, the direct wattmeter method is specified.[3,4] The British standards, issued in 1950, permit either method.[5] This subject has also been under consideration by a committee of the International Electrotechnical Commission which has drafted international standards for mercury-arc rectifiers.

One of the operating conditions influencing the arc-drop voltage is the wave shape of the anode current, its amplitude and duration. The test conditions are specified in the standards. To obtain the desired anode current wave shape, the arc-drop tests are usually made while the rectifier is operating at a d-c output voltage in the range of 250 to 600 volts or higher. A problem encountered with both test methods is the high anode-to-cathode voltage during the inverse (nonconducting) part of the cycle. For a 600-volt rectifier, for example, this voltage attains values of about 1,400 volts, while the arc-drop voltage is usually in the range of 15 to 25 volts. To permit the use of a measuring element of the desired sensitivity at the arc-drop voltage, the element is by-passed by an auxiliary rectifier during the inverse period to avoid damage from the high inverse voltage. The voltage is absorbed by an external resistance connected in series with the measuring element.

For the direct wattmeter method specified in the Swedish standards, the

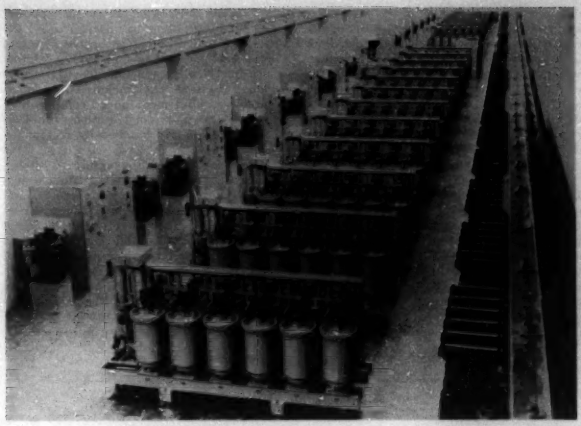


Fig. 1. View of the Wenatchee rectifier installation showing a group of excitron rectifiers for one of the potlines

wattmeter current element is supplied from the secondary of a current transformer in the anode circuit, in series with an auxiliary rectifier. The wattmeter thus receives a unidirectional current which is substantially a replica of the anode current reduced in scale of the current-transformer ratio. The arc-drop voltage is the ratio of the measured watts to the average direct current in the wattmeter current circuit.

The arc-drop tests described in the following were made on an excitron mercury-arc rectifier at the Wenatchee (Wash.) Works of the Aluminum Company of America, in June 1953.

Wenatchee Arc-Drop Tests

The Wenatchee Works has in operation four aluminum reduction potlines. Two of the potlines are supplied with d-c power by excitron rectifiers and two by ignitron rectifiers. The conversion equipment for each potline consists of six units, each with an output capacity of 12,000 amperes at 750 volts d-c. Each unit consists of a rectifier transformer and two 12-tube rectifiers. The rectifier transformer has two 6-phase double-wye d-c windings, and two tubes are connected in parallel to each phase through an anode paralleling reactor. The six units of each potline constitute a 36-phase rectifier system. A view of a group of rectifiers for one of the potlines is shown in Fig. 1.

One purpose of the arc-drop tests was to compare the oscillograph and direct-wattmeter test methods. Measurements were made on two tubes of one of the rectifiers, one tube on each side of the interphase transformer.

Test Circuit

A connection diagram of the test circuit is shown in Fig. 2. The oscillograph elements are designated by *V1* to *V5*. Element *V1* was connected to a potentiometer circuit between the anode and cathode terminals and recorded the full anode-to-cathode voltage. Element *V2* recorded the arc-drop voltage; it was connected between the anode and cathode terminals in series with resistor *R4* and was by-passed by protective thyratron tubes *TR* and *TF*. Tube *TR* was connected to fire when the anode-to-cathode voltage is negative and served to protect the oscillograph element against overcurrent from the high inverse voltage. Tube *TF* was connected to fire when the anode-to-cathode voltage is positive, to protect the element against high forward voltage before the start of anode conduction when the rectifier is operated with phase control. The tube was biased by the voltage of battery *B2* to keep it from firing during the anode conducting period in the range of the arc-drop voltage. The other oscillograph elements were used for recording the anode current.

The wattmeter was Weston model *310*, form *2* (low-power-factor type). This type was used to obtain good sensitivity. The wattmeter had 2.5/5-ampere current connections and 75/150-volt voltage terminals. A 37.5-volt terminal was brought out and used during the tests. The wattmeter voltage element was protected against overvoltage by tubes *TR* and *TF* in the same manner as the oscillograph element *V2*.

A through-type 1,000/5 current transformer was mounted on the rectifier anode cable, and its secondary was connected to the wattmeter current coil in series with a selenium rectifier and a d-c ammeter. A resistor *R5* of 20 ohms was connected across the current transformer secondary; it permits the flow of a small magnetizing current during the nonconducting part of the cycle, in order to discharge the magnetic flux produced in the *CT* core by the voltage of its burden. The oscillograph elements *V3* and *V4* were used to obtain records of the current wave shape in the *CT* secondary and the wattmeter. The oscillograph shunts were short-circuited during the wattmeter tests.

A 600/1 clamp-on current transformer was used in connection with oscillograph element *V5* for recording the anode

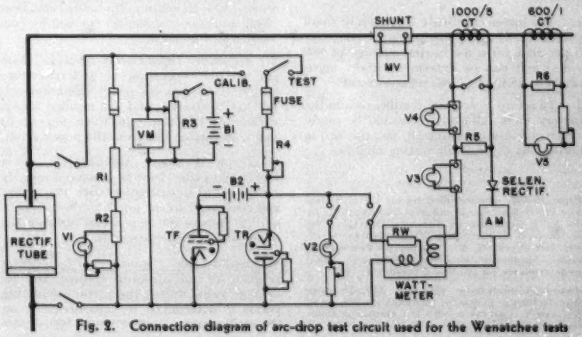


Fig. 2. Connection diagram of arc-drop test circuit used for the Wenatchee tests

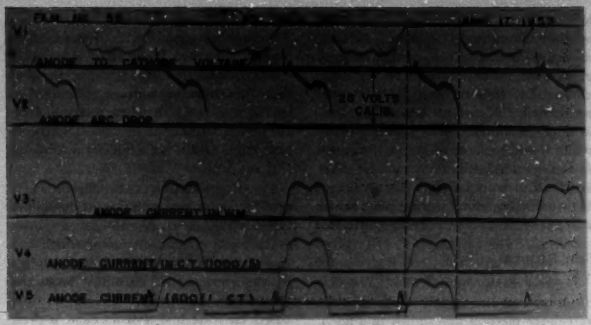


Fig. 3. Typical oscillogram taken during the arc-drop tests

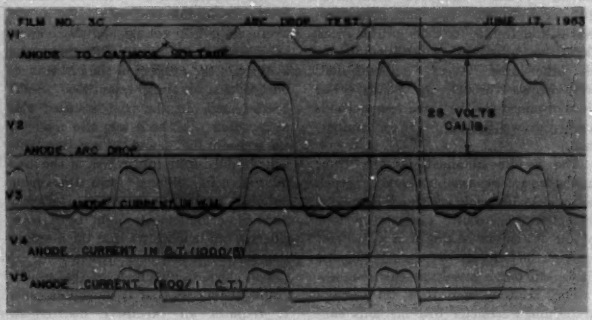


Fig. 4. Oscillogram taken during the tests, showing a trace of the arc-drop voltage obtained with a galvanometer of low natural frequency

current, as in the usual practice in oscillographic arc-drop tests. It provides a record of the anode current wave shape and is used for determining the beginning and end of the conducting period.

The average anode current was indicated by a d-c millivoltmeter connected to an ammeter shunt. Duplicate current transformers and shunts were used for the two rectifier tubes on test, and the test circuits were transferred from one tube to the other by means of double-throw switches (not shown).

Battery *B1* with rheostat *R3* and a d-c voltmeter were used for calibration purposes, and a double-throw switch was provided for transferring the arc-drop voltage circuit from the test to the calibrating connections. A calibrating voltage line was recorded on the oscillograms, and the calibrating circuit was used also in connection with the wattmeter tests.

Tests

The rectifier was operated sufficiently long at each load point tested to establish fairly stable operating conditions. Oscillograms were then made and wattmeter readings taken for both tubes on test. Because of some variations in the potline load current, there was a variation in the anode currents on successive measurements. The anode current readings obtained with the shunt and millivoltmeter in the anode circuit were used as a basis for the test results. The tests reported in this paper were made while the rectifier was operating with zero phase control which is in accordance with the ASA *C-34* standards. When oscillograms were taken the wattmeter voltage circuit was open; when wattmeter readings were taken the oscillograph circuits were open.

The oscillograms were made with a magazine film holder. A typical oscillogram is shown in Fig. 3. The traces

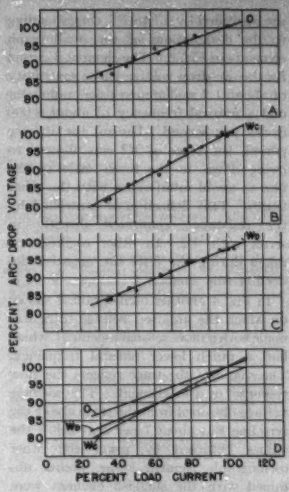


Fig. 5. Results of arc-drop tests obtained by the three test methods

A—Oscillograph method
B—Wattmeter method by calculation
C—Wattmeter method; direct voltage reading by calibration
D—Comparison

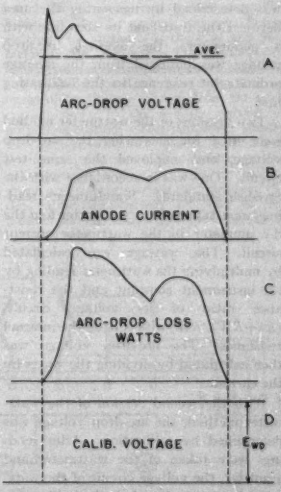


Fig. 6. Arc-drop voltage, anode current, and arc-drop loss, derived from the oscillogram of Fig. 3

are identified by *V1* to *V5*, corresponding to the designations of the oscillograph elements in Fig. 2. The element *V2* used for the arc-drop voltage had a sensitivity of about 1.2 milliamperes per millimeter and a natural frequency of about 3,000 cycles per second. During the inverse period, the trace of *V2* was allowed to swing down to the lower edge of the film to avoid interference with the other traces. This deflection is controlled by the resistance connected directly in series with *V2*. In other oscillograms taken during the tests the negative deflection was smaller.

In Fig. 4 another oscillogram is shown on which the arc-drop voltage trace was made with a more sensitive element, which had a much lower natural frequency. The shape of the trace reflects the slow response of the element. The average voltage obtained from the trace above the zero line was about 7 per cent below the values obtained by using the other oscillograph element. The points obtained with the sluggish element were not included in the test results. The oscillogram is shown for comparison with that of Fig. 3.

For determination of the arc-drop voltage, prints were made of the oscillograms and enlarged to double size. The anode conducting period was marked by vertical lines, as indicated in Fig. 3; the average ordinate of the arc-drop trace was determined by measuring the area between the trace and its zero line with a planimeter; the average arc-drop voltage was obtained from the average ordinate by reference to the calibrating line.

Two versions of the wattmeter method were used for measuring the arc-drop voltage, and employed the same test circuit. One was in accordance with the Swedish standards. Simultaneous readings were taken of the wattmeter and the d-c ammeter in the wattmeter current circuit. The wattage was calculated by multiplying the wattmeter reading by the instrument constant and the resistance ratio of its voltage circuit, $(R4+RW)/RW$, with RW as the internal resistance. The arc-drop voltage was then calculated by dividing the watts by the current.

With the second version of the wattmeter method, the arc-drop voltage was determined by calibration. After readings were taken of the wattmeter and ammeter, the voltage circuit of the wattmeter was transferred to the calibrating position and the calibrating voltage was adjusted to give the same wattmeter indication for the same current. The arc-drop voltage was indicated directly by the calibrating voltmeter. Because of variations in the load current, several trials were usually needed to obtain a reading by this method. This would not occur if the current were held stable, as during factory tests. The measurements would be facilitated also if a push-button-operated transfer contactor were used, which would reduce the swing of the wattmeter pointer during the transfer between the test and calibrating positions.

Test Results

The test results obtained by the oscillograph method and the two versions of the wattmeter method are shown in Fig. 5, on a relative percentage basis, using as a reference base the arc-drop voltage obtained by the oscillograph method at 100-per-cent load. The black points represent the results for one of the rectifier tubes tested, the circled points for the other tube. Average straight lines were drawn through the points. The lines obtained for the three test methods were redrawn in Fig. 5(D) for comparison.

The results obtained by the three methods are reasonably close to each other. At higher loads the difference is only 2 to 3 per cent. The steeper slope of line W_c might be due to a wattmeter error or the influence of external magnetic fields, but this is merely speculation.

The relative merits of the oscillograph and wattmeter test methods cannot be evaluated on the basis of the test results only. Experience gained in using these methods and other factors have to be considered, and these will be discussed in the following.

Discussion of Test Methods

Before the relative merits of the arc-drop test methods used during the Wenatchee tests are discussed, the basic principles of these methods will be reviewed briefly.

In Figs. 6(A) and (B) are reproduced traces of the arc-drop voltage and anode current from the oscillogram of Fig. 3. Fig. 6(C) shows a trace of the arc-drop loss in watts, obtained by multiplying the instantaneous values of arc-drop voltage and anode current.

By the oscillograph method as specified in the ASA *C-34* standards, the arc-drop voltage is the average of the arc-drop voltage trace during the anode conducting period, as indicated in Fig. 6(A). With the wattmeter method used in accordance with the Swedish standards, the average arc-drop loss in watts is measured and the arc-drop voltage is obtained by dividing the average watts by the average anode current. With the wattmeter calibration method, the arc-drop is equal to a d-c calibrating voltage, such as E_{WD} in Fig. 6(D), which when multiplied by the anode current will produce a watt-loss area equal to the area under the trace of Fig. 6(C).

For determination of the efficiency of a rectifier unit, the actual watt loss in a rectifier tube is required. The arc-drop voltage is a convenient factor which will give the correct watt loss when multiplied by the average tube current. From the theoretical viewpoint, a test method based on watt loss measurement is preferable to one based on voltage measurement.

The oscillograph test method is based on the assumption that the product of the average arc-drop voltage and the average anode current is substantially equal to the average of the product of the instantaneous values of these quantities. Some checks made in the past justified this assumption. A check was also made of the traces in Fig. 6, and the arc-drop voltage obtained from integration of the area in Fig. 6(C) differed from the value obtained in Fig. 6(A) by 1.5 per cent. Such close agreement might be a coincidence, but it indicates that the assumption has empirical justification.

The chief disadvantage of the oscillograph method is that it is cumbersome. It requires an expensive instrument, photographic facilities, and above-average skill and care. Determination of the test results by integrating oscillogram traces is a laborious process, and usually considerable time elapses before the test results are obtained. The advantage of the wattmeter method over the oscillograph method from the viewpoint of mechanics can be appreciated best by making comparative tests with both methods, as was done in Wenatchee.

Unless care and good judgment are used, an appreciable error could be introduced in measurements with the oscillograph method. The effect of the characteristics of the oscillograph element was pointed out previously in connection with the oscillogram of Fig. 4. There could also be variations in the arc-drop voltage traces from cycle to cycle, and several cycles picked at random may not represent a good average. Errors from this source are usually averaged out if enough points are taken for the arc-drop curve. The specifications for the oscillograph method in the *C-34* standard would have to be made more detailed to guard against such errors.

An important advantage of the wattmeter method, in addition to its firmer theoretical basis and ease of making measurements, is that it permits the use of indicating instruments for measuring the losses in rectifiers as is done for other electric equipment. More test points can be taken and the results can be plotted while the test is in progress.

The wattmeter calibration method, which is a modification of the wattmeter method specified in the Swedish standards, has several advantages over the latter. The arc-drop voltage is read directly on a voltmeter. The effect of a wattmeter error is minimized because the wattmeter reading is used only for reference. No account has to be taken of the wattmeter multiplier or the resistance ratio multiplier. The calibration method, however, cannot be used conveniently if the load current is varying. Suggestions have been made for a null calibration method, in which the calibrating and arc-drop voltages are applied to the wattmeter simultaneously in such a way that the calibrating voltage bucks the effect of the arc-drop voltage. The calibrating voltage would then be adjusted to obtain a zero reading on the wattmeter. This would facilitate testing in cases where there are moderate variations in the load current.

For the wattmeter test method some care has to be taken in the selection of circuit components such as the current transformer, auxiliary rectifier, and wattmeter. The characteristics of the circuit components and their effects on the accuracy of measurements have been dealt with by Dr. Lamm.[4]

On the basis of the experience obtained during the Wenatchee tests and for other reasons outlined in the preceding discussion, the authors believe that the wattmeter method is superior to the oscillograph method for measuring the arc-drop voltage of rectifiers and suggest that some form of the wattmeter method be considered for adoption as the standard test method. Additional tests will be required to determine which form of the wattmeter method should be adopted, what the test procedure should be, and how the circuit components should be specified.

References

1. American Standard for Pool Cathode Mercury-Arc Power Converters. *ASA C34.1-1949*, American Standards Association, New York, N. Y., 1949.

2. The Losses in Mercury-Arc Power Rectifiers and Various Methods of Measuring Them, A. Leuthold, S. Widmer. *Brown Boveri Review*, Baden, Switzerland, vol. 18, Dec. 1931, pp. 363–70.

3. Standards for Mercury-Arc Converters. *Swedish Electrotechnical Standards SEN 21-1941E*, Swedish Standards Association, Stockholm, Sweden, 1941.

4. Ein Verfahren zur Messung der Lichtbogenspannung von Mutatoren, A. Und Lamm. *Bulletin*, Schweizerischer Elektrotechnischer Verein, Zurich, Switzerland, vol. 30, no. 9, Apr. 1939, pp. 229–33.

5. Mercury-Arc Rectifier Equipments. *British Standards 1698: 1950*, British Standards Institution, London, England, 1950.

Discussion

I. K. Dortort (I-T-E Circuit Breaker Company, Philadelphia, Pa.): The authors' report of the Wenatchee arc-drop tests is an interesting corroboration of three widely accepted methods and gives some weight to their recommendation for adoption of the wattmeter method as the standard. This recommendation is based on convenience, economy, and scientific accuracy. The last item certainly has not been verified.

The implied theoretical superiority of the wattmeter method over voltage measurements can be rationalized, but is not supported in the paper. Comparison of the measured arc drops in Fig. 5 of the paper shows little difference between the three methods used during the Wenatchee tests and offers no criterion of accuracy to establish the superiority of any one of the three methods. Even theoretically, wattmeter readings are not reproducible except under conditions of duplicate a-c system and d-c load characteristics.

All three methods tested at Wenatchee are subject to indeterminate inaccuracies, are hard to apply in the factory and in the field, and require special test equipment. Therefore, before the recommendation is acted on, I would like to submit again the direct voltmeter method I proposed at the first presentation of this paper.

As first conceived and tried many years ago, the direct voltmeter method required special apparatus of considerable size and cost. Today, however, it need not be very large or expensive. Use of modern germanium and silicon rectifiers instead of vacuum tubes now makes possible the design of smaller and lighter apparatus more constant in performance, and the simplification of calibrating circuits. The test can be carried out very easily under normal operating conditions. In its simplest and most reliable form it does require operation of the rectifier at zero phase control.

Fundamentally, this method employs a metering rectifier to establish a reference direct voltage rigidly tied to the anode voltage of the rectifier under test. All voltage drops between the power source and anodes are common to both the main rectifier and the metering rectifier and therefore cancel out when the voltage difference between the two "cathodes" is measured. Only the voltage difference is measured, avoiding the error inherent in obtaining the difference between two large quantities. The difference so measured plus the calibrated drop in the metering rectifier is equal to the arc drop of the main rectifier.

Fig. 7 shows a suitable means of measuring the arc drop of a 6-phase double-wye rectifier. The principal requirements of the measuring equipment are the ability of the metering rectifier to withstand the inverse voltage peaks, its amenability to accurate determination of the forward drop, and practical means for maintaining this drop constant.

In the initial tests of this method the metering rectifier consisted of six mercury vapor diodes in a temperature-controlled housing. The d-c output was maintained constant by means of a rheostat and ammeter in the d-c lead. The loading was chosen from a d-c arc-drop curve to give a substantially constant arc drop over a range of continuous current from the peak current down to the average current of normal operation. This simplified greatly the requirements of check calibrations during the test run. The validity of the calibration was checked with an oscillograph and a rotary switch that switched the d-c calibrating current on and off for periods having ratios corresponding to 3-phase operation.

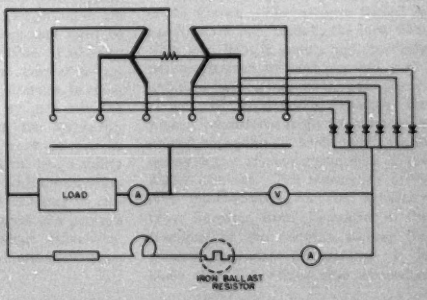


Fig. 7. Equipment for measurement of arc drop of 6-phase double-wye rectifier

Tests were conducted in 1939 on a multi-anode rectifier during bake-out. Variations of the vacuum meter were consistently and quickly followed by the arc-drop voltmeter. Slight variations in the anode plate temperature were also noted immediately on the arc-drop voltmeter. After completion of the bake-out, arc-drop readings at various currents were reproducible within 0.1 volt. Good agreement was observed between these measurements and conventional arc-drop measurements, although the comparison was not carried out with the same care and exactness as at Wenatchee.

The forward drop of modern germanium or silicon diodes and their reverse current is so small that they are ideally suited to this type of instrumentation. Since their forward drop is independent of time for short periods, calibration by direct current is valid and easy. Maintaining a reasonably constant output as with an iron ballast resistor tube gives them an essentially constant voltage drop and the equipment can therefore be simple and small.

It is not claimed that the direct voltmeter measurement is as rigorous as some of the other methods but it should be noted that the wattmeter method uses metering rectifiers in the sensitive current transformer circuit. The errors in the current transformer, flux resetting resistor, metering rectifier, and wattmeter cannot be neglected. The advantages of a continuous reading arc-drop voltmeter averaging 3, 6, or 12 anodes simultaneously, over any of the methods described in the paper, seem obvious, for both simplicity and ease of operation.

H. Winograd and **W. E. Lawton:** The arc-drop test method proposed by Mr. Dortort, using a germanium or silicon calibrated "metering" rectifier, appears to have the merit of circuit simplicity. However, a casual analysis of its application indicates that it is not as simple as it appears and it has some serious pitfalls.

For a 750-volt rectifier unit, such as those in the Wenatchee Works, the peak inverse voltage at no load is about 2,000 volts. Assuming that germanium or silicon rectifier cells are used with a rated peak inverse voltage of 200 volts, 10 cells would have to be connected in series on each phase of the auxiliary rectifier, or a total of 60 cells for a 6-phase circuit. Perhaps more would have to be used to provide a safety margin for voltage surges on the rectifier transformer. When germanium or silicon cells are connected in series, equalizer resistances have to be connected across the individual cells to distribute the inverse voltage among them. Besides the added complication, this increases the inverse current of the circuit and is likely to be a source of error. Connecting a small auxiliary rectifier to a high-power rectifier transformer without current-limiting reactors or resistors is hazardous. A breakdown in the rectifier cells of one phase is equivalent to an arc back, resulting in a high fault current. The auxiliary rectifier would have to be connected to the transformer through fuses capable of interrupting the fault at the peak inverse operating voltage.

The forward voltage drop of germanium and silicon rectifiers is affected by the junction temperature. In one commercial type of germanium cell, for example, the voltage drop is reduced by about 0.1 volt when the junction temperature is increased from 30 to 100 degrees centigrade. This would be multiplied by the number of cells in series. The temperature is influenced by such variable factors as ambient temperature, inverse voltage, and time. An unsuspected breakdown of individual cells would upset the calibration. Any delay in the firing point of the main rectifier anodes, which might be due to a high pickup voltage could also introduce an error. The wave shape of the phase currents in the metering rectifier would also influence the voltage drop. These are some of the concealed sources of error which would make this method unreliable.

Mr. Dortort stated that the recommendation of the wattmeter method is based on the claim of scientific accuracy, which has not been verified. No such claim was made in the paper. Scientific accuracy can hardly be expected in a field test. Our recommendation of the wattmeter method was based on the fact that it is simpler and more reliable than the oscillograph method; that it also has a firmer theoretical basis from the viewpoint of rectifier efficiency determination, because the power loss is measured rather than the average voltage. The characteristics of the current transformer and other components of the wattmeter test circuit would have to be specified in the standards to obtain the desired accuracy, as is the usual practice in engineering tests.

Let-Go Currents and Voltages

C. F. DALZIEL
MEMBER AIEE

F. P. MASSOGLIA
ASSOCIATE MEMBER AIEE

Synopsis: Let-go current curves are presented for commercial 60-cycle a-c and for direct current. New data are presented which confirm previously published let-go current curves for direct currents having various amounts of a-c ripple content, together with new data for 60-cycle a-c let-go voltages. The voltage analysis is extended to cover direct current, thus rounding out the subject of let-go hazards. Reasonably safe let-go currents and voltages are computed from tests made on substantial numbers of normal, healthy individuals. It is believed that reasonably safe limits may be based on the predicted response for 99½ per cent of a large group. The probability of receiving dangerous electric shocks must be anticipated, and preventive measures must be considered vital ingredients in the design of machines, in the framing of operating instructions, and in work procedures.

FOR many practical purposes the maximum electric current safe for man is just a little less than that causing him to "freeze" to a circuit. This current is called his let-go current; this is the maximum current he can tolerate and still be able to release or let go his grasp of an energized conductor by using the muscles directly stimulated by that current. Carefully conducted experiments involving several hundred volunteer subjects during the past 17 years have resulted in a substantial amount of data and in methods of analysis which permit predictions to be made for many different conditions with an accuracy believed sufficient for many engineering purposes. Although there was considerable variation among different individuals, and minor changes were made in experimental procedures from time to time to improve laboratory techniques, the muscular reactions to electric current, while varying somewhat in degree, gave results which were in very substantial agreement. As long as the current pathway involves the hand and forearm, the muscular contractions, discomfort, and pain increase as the current is increased, and currents of only a few milliamperes are sufficient to prevent a subject from releasing an electrode grasped in the palm of the hand.

Without exception all let-go current determinations, with the current pathway involving the chest to the other hand or to the feet, were conducted in the presence of a physician in co-operation with the University of California Medical School. The subjects were given physical examinations including electrocardiograms. Only those over 18 years of age, in good physical condition, who had not recent illness, and who had normal blood pressures and electrocardiograms were used. Only contact areas with healthy skin were subjected to the tests.

Experience has demonstrated that a person can withstand repeated exposure to current up to his let-go limit with no ill aftereffects, at least for the time required for him to release his grasp of an energized electrode. Although let-go values as determined in the laboratory may be considered somewhat academic, it is obvious that in accidental contact, release from currents in excess of an individual's let-go value is conjectural. A person can often let go at slightly higher values when dared, and higher values would be expected when a victim realized the seriousness of his predicament. However, contact by a perspiring workman, when standing in water, or when working inside a grounded metallic structure, closely parallels the hazardous conditions simulated in these experiments. In many accidents a victim can free himself by breaking the conductor, or his body weight may assist him in interrupting the circuit; however fortuitous circumstances must not be relied upon to provide safety for human life. Currents only slightly in excess of a person's let-go value are very painful, frightening, and hard to endure for even a short time. Failure to interrupt promptly the current is accompanied by a rapid decrease in muscular strength caused by pain and the fatigue associated with the severe involuntary muscular contractions. It would be expected that the let-go ability would decrease rapidly with the duration of contact. Prolonged exposure to currents only slightly in excess of a person's let-go limit may produce exhaustion, asphyxia, collapse, and unconsciousness followed by death.

Sixty-cycle sine-wave let-go currents determined for 134 men and 28 women are shown in Fig. 1. In these tests the subjects held and then released a test electrode consisting of a no. 6 or no. 7 polished copper wire. The circuit was completed by placing the other hand or foot on a flat brass plate, or by clamping a conductive band lined with saline-soaked gauze on the upper arm. After one or two preliminary trials to accustom the subject to the sensations and muscular contractions produced by the current, the current was increased to a certain value and the subject was commanded to let go of the wire. If he succeeded, the test was repeated at a current of a slightly higher value. If he failed, a lower current was used, and the values were again increased until the subject could no longer release the test electrode. The end point was checked by several trials, and the highest value was taken as the individual's let-go value so as to eliminate the effects of fatigue. The experimental points plotted in the figure were obtained at the instant of releasing the electrode, with the subjects' hands wet with salt-water solution to secure uniform conditions and to reduce the sensation of burning caused by high current densities at tender spots. Other tests were made with dry hands, hands moist from perspiration, and hands dripping wet from weak acid solutions. The effect of the size of the electrodes was also investigated. It was found that the location of the indifferent electrode, the moisture conditions at the points of contact, and the size of the electrodes had no appreciable effect on the individual's let-go current. It is believed that the results obtained from tests, in which hands, wet with saline solution grasp and then release the small copper wire, may be used to predict let-go currents of a specified degree of safety for normal men and women with an accuracy sufficient for many practical purposes.

Sixty-cycle let-go currents were tested on 28 women. The subjects ranged in age from the late 'teens to the early 20's. They were light in stature and obviously not accustomed to hard physical work, and their forearm muscles were not particularly well developed. Although the women volunteered freely for the test, it proved impossible to develop enthusiasm or any degree of competitive spirit at the higher currents. The results are probably representative for the sedentary type; however, from observation of the reactions of the subjects having the greatest muscular development, it is likely that the values were somewhat

Paper 56-111, recommended by the AIEE Electric Welding Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Winter General Meeting, New York, N. Y., January 30-February 3, 1956. Manuscript submitted October 4, 1954; made available for printing November 21, 1955.

C. F. DALZIEL is with the University of California, Berkeley, Calif., and F. P. MASSOGLIA is with the San Francisco Naval Shipyard, San Francisco, Calif.

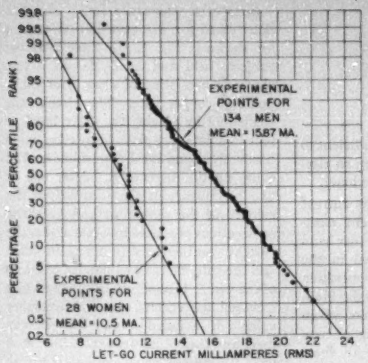

PERCENTAGE (PERCENTILE RANK)
EXPERIMENTAL POINTS FOR 134 MEN MEAN = 15.87 MA.
EXPERIMENTAL POINTS FOR 28 WOMEN MEAN = 10.5 MA.
LET-GO CURRENT MILLIAMPERES (RMS)

Fig. 1 (left). 60-cycle let-go current distribution curves for men and women

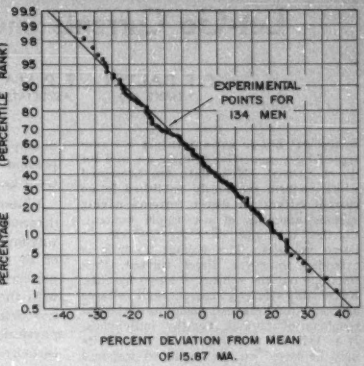


Fig. 2 (right). 60-cycle let-go current deviation curve

lower than those which would have been obtained had a group of women accustomed to physical labor been used. Results based on these data, therefore, should be conservative and on the side of safety. These data constitute the available let-go data on women.

The mean or average value, which is called the let-go threshold, was established at 15.87 and 10.5 milliamperes (ma) for men and women respectively. Probably women's threshold let-go current is only 66 per cent (%) of the let-go threshold value found for men because their nervous systems are more sensitive and their muscular development is poorer, and not because there is any difference in sex.

There was considerable variation in the individual's let-go current in repeated tests made at weekly intervals, the trend usually being toward slightly higher values; therefore, the largest current released on the first test was taken as the individual's let-go current. This was done to include the element of surprise to as great an extent as possible and to give conservative results. Psychological factors, especially fear and competitive spirit, were the most important causes for the variations. Physiological factors played an important part but so far their exact mechanism remains unknown. It seemed that the let-go current in both sexes was related to the muscular development of the wrist and forearm. Husky subjects having low let-go values could almost invariably be persuaded to continue the test until their values were in line with others of similar physique; however, attempts to correlate let-go currents with physical measurements of the forearm, wrist, strength of grip, etc., were inconclusive.

Reasonably Safe Currents

The principle of biological variability is so universally recognized that no attempt should be made to specify any electric current as safe for all people. The press contains frequent accounts of fatalities ascribed to heart failure caused by overexcitement, intense emotion, or shock (shock of injury, not electric shock). Some of the subjects volunteering for these tests became frightened and trembled all over; some even complained of pain when holding the test electrode before the circuit was energized. Although these persons were not used in the experiments, the experiences dramatically illustrate the possibility that a person with a diseased heart might succumb from any contact, or even the fear of contact, with an electric circuit. This possibility must be recognized, and an occasional death is to be expected as a result of casual contact involving currents known to be harmless to the great majority of healthy individuals. Quite aside from determining an absolutely safe electric current for all human beings is the practical problem of determining a current which would be reasonably safe for most normal, healthy individuals.

It has been found that the response to electrical stimulation obtained from substantial numbers of subjects inevitably follows a normal distribution. This fact is illustrated in the distribution curves of Fig. 1, in which the majority of the individual points fall closely about the straight lines on the probability graph paper. It has been found that the greater the number of individual points the more closely the points fall about a straight line. The slightly greater scattering of the points about the distribution curve for the 28 women in contrast to the consistent response for 134 men illustrates the point. The distribution curve is a ready means for predicting the probable response for a given percentage of a large group. It is believed that data have been obtained from sufficiently large groups to permit valid predictions to be made, not only for these particular groups but also for a large segment of the normal, healthy population of the country.

Without a single exception experimental points representing the very low currents diverge somewhat from the distribution curves. This is consistent with the observation that every normal person should be able to release some current. In contrast, because of the nature of the probability scale, the distribution curves yield a theoretical probability for zero current. A great deal of thought has been given to establishing a reasonable criterion for safety and an arbitrary limit based on the theoretical response for 99½% of a large group has been quite generally accepted by the profession.[1-8] Thus, it is concluded that the reasonably safe 60-cycle let-go current for men and women is approximately 9 and 6 ma respectively.

An alternate method of plotting the data is illustrated in Fig. 2, in which the points for the 134 men are plotted as percent deviations from the mean of the group. This method has the advantage of increased accuracy in analyzing the data, as it has been found that as long as the sensations and other manifestations resulting from electrical stimulation are similar, the deviation curves will have the same or nearly identical slopes. In contrast, the corresponding distribution curves generally have different slopes.

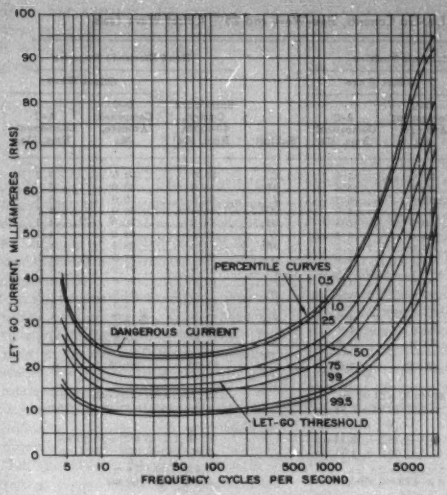


Fig. 3 (left). Effect of frequency on let-go current for men. Current values becoms dangerous progressively to an increasing number of persons, as indicated by the percentile values on the right-hand side of the curves. Values for women are approximately 66% of the current values shown on the curves

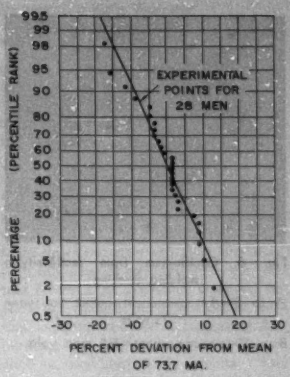


Fig. 4 (right). D-c release current deviation curve

Such tests are time-consuming, and smaller groups were used to investigate other conditions. For example, in so far as could be ascertained, considering limitations of accuracy of the data and the relatively limited numbers of subjects used, the 60-cycle let-go values for 134 men, for 28 women, and for groups varying from 25 to 30 men, with the use of sine-wave alternating currents from 5 to 5,000 cycles, yielded deviation curves all of which had the same slope.

The following procedure is believed to yield an accuracy for small numbers of subjects comparable to that for the 134-man group. Comparative let-go tests are made on 60-cycle sine-wave alternating current with the use of the copper-wire electrode and a selected test condition, such as an irregular wave form, higher frequency, etc. The order of the tests is alternated among the subjects to equalize effects of fatigue. The distribution curve is plotted for the selected test condition, and wild points are discarded. Per-cent deviations are computed for the points retained, and the deviation curve is plotted and compared with Fig. 2. The data are considered reliable if the response approximates a normal distribution, and if the slope of the resulting deviation curve is consistent with the reported physiological reactions. It may be of interest that only three points were discarded in reducing the data for the several tests mentioned in the foregoing. The projected mean for a large sample is computed as follows:

$$\text{mean or threshold value} = \frac{\text{mean 60-cycle let-go current for 134 men}}{\text{mean 60-cycle let-go current for sample}} \times$$

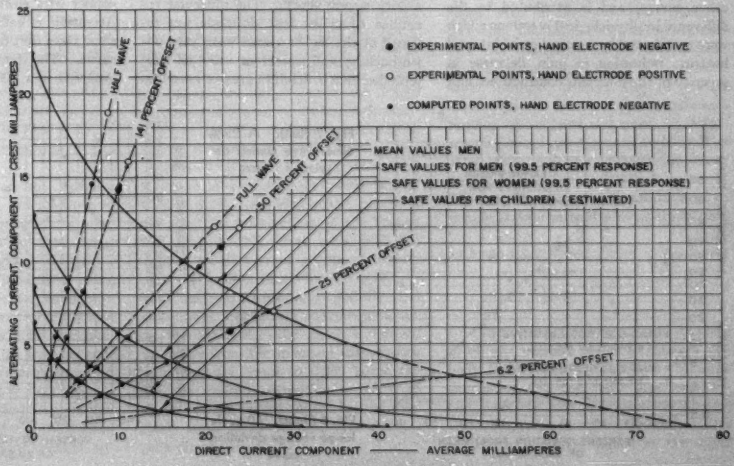

ALTERNATING CURRENT COMPONENT — CREST MILLIAMPERES
DIRECT CURRENT COMPONENT — AVERAGE MILLIAMPERES
EXPERIMENTAL POINTS, HAND ELECTRODE NEGATIVE
EXPERIMENTAL POINTS, HAND ELECTRODE POSITIVE
COMPUTED POINTS, HAND ELECTRODE NEGATIVE
MEAN VALUES MEN
SAFE VALUES FOR MEN (99.5 PERCENT RESPONSE)
SAFE VALUES FOR WOMEN (99.5 PERCENT RESPONSE)
SAFE VALUES FOR CHILDREN (ESTIMATED)
HALF WAVE
FULL WAVE
D-C
40 PERCENT OFFSET
50 PERCENT OFFSET
25 PERCENT OFFSET
6.2 PERCENT OFFSET

Fig. 5. Effect of wave form on let-go currents for men, women, and children

Fig. 6. Oscillogram from rectifier-type welding machine

A to B = d-c component output voltage = 75 volts

B to C = a-c component = 5.5 crest volts

Table I. Let-Go Currents, Sine-Wave and 25% Offset Wave Tests

	Hand Electrode Positive				Hand Electrode Negative		
		Offset Wave				Offset Wave	
Subject	Sine-Wave Control, 60-Cycle Rms, Ma	D-C Component Average, Ma	A-C Component Rms, Ma	Subject	Sine-Wave Control, 60-Cycle Rms, Ma	D-C Component Average, Ma	A-C Component Rms, Ma
P	18.8	36.0	5.6	*C*	16.5	30.0	5.2
C	16.5	34.0	5.6	*D*	14.0	30.0	5.3
B	15.0	32.0	5.7	*I*	16.2	30.0	5.6
I	16.2	31.0	5.9	*A*	16.0	26.0	4.7
M	14.5	30.0	5.3	*M*	14.5	26.0	4.6
D	14.0	30.0	5.2	*G*	12.4	24.0	4.3
A	16.0	30.0	5.5	*B*	15.0	23.5	4.2
G	12.4	27.5	4.9	*N*	12.5	23.0	4.1
E	12.8	25.0	4.4	*F*	11.0	21.2	3.9
H	13.3	25.0	4.4	*E*	12.8	20.0	3.7
K	12.0	25.0	4.6	*H*	13.3	20.0	3.5
O	12.0	24.0	4.3	*O*	12.0	18.0	3.2
F	11.0	23.0	4.2	*J*	11.0	18.0	3.4
N	12.5	23.0	4.1	*L*	11.1	17.5	3.3
L	11.1	23.0	4.3	*K*	12.0	17.0	3.0
Mean	13.87	27.90	4.93		13.35	22.95	4.13
Corrected mean:							
D-c average	$27.90 \times 15.87/13.87 = 31.92$ ma				$22.95 \times 15.87/13.35 = 27.28$ ma		
A-c crest	$= \sqrt{2} \times 4.93 \times 15.87/13.87 = 7.98$ ma				$\sqrt{2} \times 4.13 \times 15.87/13.35 = 6.94$ ma		
99½ percentile:							
D-c average men	$= 31.92\ (1.0 - 0.432) = 18.13$ ma				$27.28\ (1.0 - 0.432) = 15.50$ ma		
A-c crest men	$= 7.98\ (1.0 - 0.432) = 4.53$ ma				$6.94\ (1.0 - 0.432) = 3.94$ ma		
D-c average children	$= 0.5 \times 18.13 = 9.1$ ma				$0.5 \times 15.50 = 7.8$ ma		
A-c crest children	$= 0.5 \times 4.53 = 2.3$ ma				$0.5 \times 3.94 = 2.0$ ma		
D-c average women	$= 18.13 \times 10.5/15.87 = 12.0$ ma				$15.50 \times 10.5/15.87 = 10.2$ ma		
A-c crest women	$= 4.53 \times 10.5/15.87 = 3.0$ ma				$3.94 \times 10.5/15.87 = 2.6$ ma		
% ripple	$7.98/31.92 = 25.1\%$				$6.94/27.28 = 25.4\%$		
Polarity effect	$27.28/31.92 = 85.5\%$						

mean of sample for given test condition (1)

The current required for other percentile ranks is obtained by using the following equation

current = mean of sample (1 ± deviation from mean corresponding to percentile rank desired) (2)

The effect of frequency on let-go currents for frequencies from 5 to 10,000 cycles is shown in Fig. 3. The tests were conducted on groups consisting of 25 to 30 men with the use of the copper-wire electrode with sine-wave shapes maintained. The data were analyzed as previously discussed and, except for the 10,000-cycle test, all the deviation curves had identical slopes. The slope of the deviation curve for the highest frequency was considerably greater than that for the other curves, and is attributed to the difference in physiological reactions which were manifested by a sensation of internal heating, reduction in pain, decrease in severity of the muscular contractions, and in a material increase in the time required to release the conductor.

Tests on steady or gradually increasing direct current produced sensations of internal heating rather than muscular contractions. Sudden changes in current produced severe muscular contractions, and interruption of the current produced a very severe shock. The muscular reactions when the test electrode was released at the higher values were very objectionable, and sooner or later all subjects declined to attempt greater currents. The maximum a subject would withstand and then release was termed his release current. This value of current represents the limit of endurance rather than the let-go limit, and it is a psychological rather than a physiological limit. The deviation curve resulting from tests on 28 men in which the hand grasped the copper wire electrode is given in Fig. 4. Attention is called to its much greater slope than that for the 60-cycle deviation curve. The data for the d-c tests were analyzed in a manner similar to that for

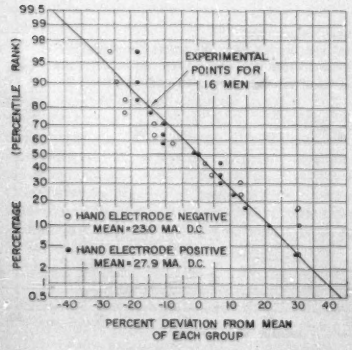


Fig. 7 (left). A 25% offset wave current deviation curve

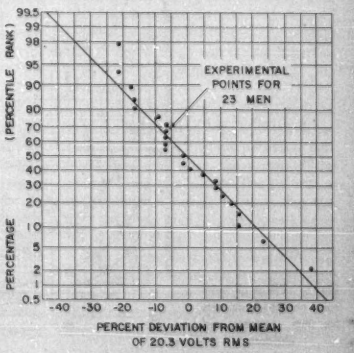


Fig. 8 (right). 60 cycle let-go voltage deviation curve

alternating currents; for example, the mean or threshold value for direct current is determined by using equation 1, thus giving

$$\text{d-c threshold current} = \frac{15.87}{15.37} \times 73.7 = 76.1 \text{ ma d-c} \quad (3)$$

The data were obtained with a wire electrode of negative polarity, as it was found that the muscular reactions were more objectionable than when releasing an electrode of positive polarity; hence, the threshold as determined is considered conservative.

A d-c release current of 56 ma was obtained by one woman at the time the tests were being made on the men. The tests on the other 27 women were concluded by having them release a maximum of 35 ma negative polarity, which each of them did without complaint or difficulty, but lack of time prevented determining limiting values. Assuming that the 60-cycle let-go current ratio of women-to-men equals 66% applies, then the d-c release threshold for women is 50.4 ma.

The effect of wave form on let-go currents for composite waves consisting of 60-cycle and direct components is shown in Fig. 5. The mean let-go values for the various composite wave forms follow a smooth curve if the direct component is plotted on one axis versus the a-c peak in the forward direction on the other axis. The curve is based on data obtained with the use of 60-cycle sine waves, rectified half-waves, rectified full waves, 50% offset waves, 141% offset waves, and pure direct current.[3] The curve representing the mean or threshold response was drawn through the computed mean values for the various tests with the hand electrode of negative polarity to give conservative results. The meaning of the phrase "a-c peak in the forward direction" is illustrated in Fig. 6. This will be discussed in the section which follows.

Rather small numbers of subjects were used for most of the composite-wave let-go tests, and insufficient data were obtained between the 50% offset waves and the d-c tests. Points obtained on 16 men during the spring of 1953, using a 25% offset wave with the hand electrode both positive and negative are also included in Fig. 5. Except for one man, the 1953 subjects were an entirely new group, and it is somewhat surprising that the computed points fall so very close to the original curves published 11 years ago. The polarity effect was also in substantial agreement with the original findings, being 85.5% versus 87.5%. Curves representing reasonably safe electric currents for men were computed with the use of the 99½% criterion; values for women were taken at 66%, and the curves for children were arbitrarily assumed at 50% of the corresponding values for men, for the reasons given in reference 2.

Data for the 25% offset wave were measured with a d-c milliammeter and the a-c component was obtained from a peak responsive General Radio type-*1803A* vacuum-tube voltmeter connected in shunt to a 1,000-ohm noninductive resistor. The instruments were connected in a series circuit comprising a variac and 90-volt battery with potentiometer control, reversing switch, and the test and indifferent electrodes. It was believed that the d-c readings were the most reliable, and the experimental points for the direct component for both polarities are plotted in Fig. 7. Although there is some scattering of the points about the deviation curve, the data are typical of the results obtained when using small numbers of subjects. It is interesting to note that the deviation curve has the same slope as that previously found for the 134-man 60-cycle response. Table I represents both data and detailed calculations.

An evaluation of the relative let-go current hazard of a typical modern rectifier-type welding machine, with the output voltage characteristics shown in the oscillogram of Fig. 6, may illustrate application of the foregoing analysis. The d-c output voltage is represented by the distance A to B = 75 volts. The a-c peak in the forward direction is B to C and is approximately 5.5 peak volts, and comparison with the timing wave indicates that the ripple frequency is 360 cycles. The % ripple is 5.5/75 = 7.3%. It is apparent from Fig. 3 that any percentile curve shows the let-go current at 360 cycles to be approximately 118% of the value at 60 cycles; and since the a-c components of Fig. 5 have a frequency of 60 cycles, the equivalent % ripple from a hazard viewpoint is 7.3/1.18 = 6.2%. An idea of the hazard is given by the intersections of the broken line marked 6.2% offset, along with the various let-go current curves of Fig. 5. (The 6.2% offset line is located by drawing a line from the origin to a point located as follows: Select any convenient value on the abscissa. The point is located vertically from the point on the abscissa at a distance equal to 0.062 times the abscissa value.) Attention is called to the increased hazard caused by the relatively small ripple content over that for pure direct current.

Reasonably Safe Voltages

Although the deleterious effects of electric shock are caused by the current actually flowing through the human body, in accidents the voltage of the circuit is usually the only electric quantity known with certainty. While current and voltage are related by Ohm's law, it is seldom that the contact resistances with the body can be determined with any degree of accuracy. On very high-voltage circuits, skin and contact resistances break down instantly and, thus, they may play only a minor role in limiting the current received by a victim; however, on the lower voltages the resistances at contact locations become of increasing importance and are of paramount importance on very low-voltage circuits. The moisture and chemical content of the epidermis, the area actually making contact with the skin, the moisture condition and pressure at the contact locations, and the resistance of electrical burns are all unknown quantities. Obviously, wet contacts create a most dangerous condition for receiving an electric shock. From the foregoing, it is evident that the problem of estimating reasonably safe voltages is much more difficult than that of establishing reasonably safe currents. However, the conclusions are consistent with the opinions of international authorities, and it is anticipated that the results may be useful in evaluating relative hazards.

Sixty-cycle sine-wave let-go currents and voltages were recently obtained on a group of 23 subjects.[4] The let-go voltage deviation curve of Fig. 8 was obtained with the subjects firmly grasping the small wet copper-wire electrode and with the indifferent electrode consisting of a lead strip wrapped with saline-soaked gauze clamped around the associated upper arm. It is noted that a sufficient number of subjects was used to determine a normal distribution, and that the let-go voltages follows a deviation curve similar to that for the 134-man group of Fig. 2. The projected mean for a large sample is determined by using equation 1

$$\text{a-c mean let-go voltage, hand to upper arm} = \frac{15.87}{14.05} \times 20.33 = 23.0 \text{ volts rms}$$

In many cases of accidental electric shock the current pathway through the body is between the hands, or between the hands and the feet. The hand-to-feet pathway was simulated by having the subjects grasp with wet hands a pair of 6-inch long-nose pliers in the right

Table II. Pathway Correction Factor, 60-Cycle A-C

Subject	Pathway Correction Factor
L	0.959
G	0.932
O	0.888
T	0.885
F	0.864
P	0.835
J	0.823
K	0.820
Y	0.811
C	0.808
X	0.806
D	0.780
B	0.766
A	0.738
M	0.737
S	0.721
E	0.676
R	0.667
U	0.660
H	0.634
L	0.592
Mean	0.781

hand when standing barefoot in a bucket of salt water to a depth of about 4 inches. Because of the violent muscular reactions and frequent loss of balance when the subject struggled to release the test electrode, it was not safe to conduct let-go tests and, instead, a series of simultaneous current and voltage readings were taken for both indifferent electrode locations with currents from about 30 to 90% of the individual's let-go value. The tests were made on the same day, with the order alternated between different subjects so that unpredictable variations would average out. A pathway correction factor was determined for 21 subjects, and is defined as follows

pathway correction factor =

$$\frac{\text{voltage}}{\text{current}}\ \text{voltage and current with right hand holding pliers to both feet in bucket of salt water}$$

$$\frac{\text{voltage}}{\text{current}}\ \text{voltage and current with right hand holding small copper wire to armband on upper arm}$$

Values for the various subjects, including the mean, are given in Table II.

It had been anticipated that the resistance from the hand to the feet, being a longer pathway, would be higher than the resistance from the hand to the armband. Such was not the case, which indicates that the contact resistance at the armband predominated. The threshold 60-cycle let-go voltage for the hand-to-feet pathway for a large sample is obtained by multiplying the mean pathway correction factor by the a-c mean let-go voltage, or 0.781 × 23.0 = 18 volts rms. Applying the 99½% criterion, the reasonably safe 60-cycle let-go voltage is computed from equation 2

Fig. 9. Subject in process of determining his let-go current and voltage

a-c reasonably safe voltage, hand to feet = 18.0 (1 − 0.432) = 10.2 volts rms

(Deviation from the mean at 99½% = −0.432 is from Fig. 8.) Let-go and release voltages for the hand-to-hand pathway were obtained from the data given in Tables I and II of reference 3. These 60-cycle let-go and release tests were comparable, since the contact conditions consisted of grasping the small copper-wire electrode in the wet palm of the right hand, and the left hand was held firmly on a flat brass plate 8 inches in diameter and wet with salt water. The data were reduced according to the methods previously described. Sixty-cycle let-go values were available for 54 subjects, but the six highest voltage values were discarded for the response to approximate a normal distribution. Although the data did not follow the anticipated response as closely as desired, it is believed that the substantial number of subjects permits reasonably accurate computations. The 60-cycle currents often produced muscular reactions at the left hand which caused the hand and fingers to curl, and it was found necessary for a bystander to apply pressure on the back of the hand to assure good contact with the brass plate. It is likely that this difficulty is responsible for the poorer response. Fig. 9 illustrates a subject in the process of determining his let-go values. The threshold for the hand-to-hand current pathway for a large group is given by

$$\text{a-c threshold let-go voltage, hand to hand} = \frac{15.87}{14.81} \times 34.4 = 36.9 \text{ volts rms}$$

The reasonably safe 60-cycle let-go voltage for this current pathway is obtained as before

a-c reasonably safe voltage, hand to hand = 36.9 (1 − 0.432) = 21 volts rms

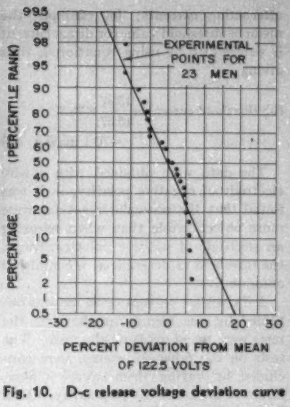


Fig. 10. D-c release voltage deviation curve

D-c release values were available for 23 subjects, and no points were rejected. The voltage deviation curve of Fig. 10 has the same slope as the corresponding d-c release current deviation curve of Fig. 4, and is offered as substantiating the analysis. The threshold for the hand-to-hand current pathway for a large group is computed with equation 1

$$\text{d-c threshold release voltage, hand to hand} = \frac{15.87}{15.21} \times 122.5 = 127.8 \text{ volts}$$

and the reasonably safe release voltage is

d-c reasonably safe voltage, hand to hand = 127.8 (1 − 0.185) = 104.2 volts

(Deviation for the mean at 99½% = −0.185 is from Fig. 10.) The ratio of the d-c mean release voltage to the corresponding a-c mean let-go voltage for the hand-to-hand pathway and the same contact conditions is 127.8/36.9 = 3.46. This permits estimating the direct voltages for the hand-to-feet current pathway as follows

d-c threshold release voltage, hand to feet = 3.46 × 18.0 = 62.3 volts d-c

and the reasonably safe release voltage is

d-c reasonably safe voltage, hand-to-feet = 62.3 (1 − 0.185) = 51 volts d-c

From the foregoing it is apparent that, for wet contact conditions, the reasonably safe 60-cycle let-go voltages for man, for the major current pathways through the body, are between about 10 and 21 volts rms, and the corresponding voltages for direct current are 51 to 104 volts.

The reader is cautioned against interpreting the differences between the

voltage values as a true index of the relative hazard between alternating current and direct current. Current is the proper criterion of electric shock intensity, and the hazard from the proposed reasonably safe voltages would be greatly increased if contact occurred at locations where the skin was lacerated or if local high current densities produced material breakdown of the skin. Currents of the let-go level are more than sufficient to produce very serious burns.

Opinions of International Authorities

H. B. Whitaker published an exhaustive study of the effects of electric shock with regard to electric fence controllers.[5] With regard to voltage he concludes; "Where no inherent current-limiting features are incorporated in the device, the maximum safe voltage to which an individual may be subjected should not exceed 12 (60-cycle a-c). This is based upon the theory that a potential of 12 volts or less will rarely, if ever, cause a breakdown of skin resistance sufficient to permit a current flow through the body of such intensity as to cause lack of muscular control or physical injury to the person."

W. B. Kouwenhoven[6] states: "People recognize that high voltages are dangerous. However, they should be equally careful of low voltages. There are a number of cases on record where contact with 60 and 65 volt circuits of commercial frequencies have resulted in fatal accidents. The lowest voltage fatality of which the author has any record occurred at 46 volts, 60 cycles. It is probable that circuits of 24 volts or less may be considered as safe under practically all conditions."

At a conference sponsored by the Comité Medical of the Électricité de France during the CIGRE meetings of June 1952, French authorities considered the maximum voltages safe for man were approximately 24 volts for 50-cycle a-c and 50 volts for direct current.

Conclusions

Because electric currents of only a few milliamperes and, with wet conditions, of only a few volts are sufficient to freeze a victim to an electric circuit, the electrical hazards inherent in electric equipment must be recognized, and measures must be taken to prevent electric shock accidents. The probability of receiving electric shocks must be anticipated, and preventive measures must be considered vital ingredients in the design of machines, in the framing of safety codes, operating instructions, and work procedures.

References

1. Effect of Frequency on Let-Go Currents, Charles F. Dalziel, Eric Ogden, Curtis E. Abbott. *AIEE Transactions* (*Electrical Engineering*), vol. 62, Dec. 1943, pp. 745–50.

2. Effect of Wave Form on Let-Go Currents, Charles F. Dalziel. *Ibid.*, pp. 739–44.

3. Electric Shock, Charles F. Dalziel, John B. Lagen, Joe L. Thurston. *Ibid.*, vol. 60, Dec. 1941, pp. 1073–79.

4. A Study of the Hazards of Open-Circuited Current Transformers, Frank P. Massoglia. *M.S. Thesis*, University of California, Berkeley, Calif., Jan. 28, 1954.

5. Electric Shock as It Pertains to the Electric Fence, H. B. Whitaker. *Bulletin of Research No. 14*, Underwriters' Laboratories, Inc., New York, N. Y., Dec. 1939.

6. Effects of Electricity on the Human Body, W. B. Kouwenhoven. *Electrical Engineering*, vol. 68, Mar. 1949, pp. 199–203.

Discussion

W. B. Kouwenhoven (The Johns Hopkins University, Baltimore, Md.): Prof. Dalziel's work on let-go current is a valuable contribution to our knowledge of the effects of electricity on the human body. In this paper the authors have attempted to extend the data to include the voltage of the circuit. For example, they find that a reasonably safe direct voltage, hand to hand, is 104.2 volts and hand to feet 51 volts.

At the Johns Hopkins Hospital, the skin resistance of the bodies of certain patients is measured as an aid to diagnosis. The resistances are measured between an indifferent electrode at some convenient location and a small silver wheel which is moved over the body surface. In this work a potential of 90 volts direct current is used and we have experienced no difficulties, and no one has been injured.

The outer layer of our skin, the epidermis, is our best protection against electric shock. The resistance of the epidermis varies widely over the surface of the body. It is low on the palms of the hands, the soles of the feet, the axillae, the face, the groin, and certain other body areas. In passing the mid-line from the lower surface of a finger to the back there is an increase in resistance of eight- to tenfold. The skin resistance also varies with the applied voltage, its conditions as to wetness, and other factors. Once the current gets inside the body the resistance is practically uniform and is about 100 ohms per centimeter[4].

In view of the wide variations in the resistance of the human skin, I believe that it is impossible to predict "safe let-go voltages" unless all of the conditions are accurately known.

A. U. Welch (General Electric Company, Fitchburg, Mass.): Contact area is important in affecting current through the body. Wet gloves can amplify shock beyond that obtained on the bare hand by increasing effective contact area.

It is interesting to note that the trend toward low voltage (40 or less) d-c power sources for the high-speed consumable electrode processes is definitely in the direction of increased safety.

D. B. Robinson (American Can Company, Maywood, Ill.): Do the let-go currents and voltages discussed in the paper cause muscular reactions such as to promote "grabbing" of contact?

K. A. Krasin (Chicago Bridge and Iron Company, Chicago, Ill.): Does not the higher voltage contacted on muscular reaction result in an opposite effect to that encountered at lower voltage?

C. F. Dalziel and **F. P. Massoglia:** The authors are in full agreement with Mr. Kouvenhoven that let-go current is a valid criterion of electric shock hazard, and that the corresponding let-go voltage is uncertain unless all of the conditions are accurately known. The maximum safe voltage when contacting a circuit of maintained voltage is largely dependent upon the dryness and dielectric strength of the epidermis at the points of contact. In the paper the let-go voltage study was confined to estimating minimum probable let-go voltages under very wet contact conditions. These conditions consisted of standing ankle deep in salt water and grasping a bare wire or a pair of pliers with the hand dripping wet with salt water. Conditions similar to this may prevail when working in very wet locations.

Perhaps the greatest value of the study is to emphasize the potential danger from very low voltages with the hope that such information will result in creating more respect for ordinary house circuits among users of electrical apparatus and appliances, including all members of the family. Fortunately, low-voltage hazards are virtually eliminated by the application of a little common sense, such as: de-energizing the circuit whenever possible; never working on a live circuit, or using electrically powered tools or appliances barehanded when standing on metal floors, in water, on wet ground, or in the rain. Instead, dry gloves and rubber overshoes should be worn; and well-grounded and wet areas should be covered with a dryboard, a couple of sacks, or an old coat.

In reply to Mr. Welch, since contact resistance is a predominating factor in limiting the current in low-voltage accidents, anything that will increase the effective area of contact with the body will increase the hazard. Wet gloves would be expected to increase greatly the potential

hazard, and leather gloves are preferable to cotton.

In reply to Mr. Robinson, alternating currents flowing through the body stimulate the nerves and muscles and produce sensations of heat. As the current is increased, the sensation of tingling and the accompanying sensation of heat increase. With further increase of current the muscular reactions increase in severity, sensations of pain develop, and voluntary control of the muscles that lie in the current pathway becomes increasingly difficult. Finally a value of current is reached for which the subject cannot control the affected muscles, and if contact is made by grasping a conductor in the hand, he finds that he cannot release his grasp of the conductor, and is said to "freeze" to the circuit.

A rather striking experiment may be safely conducted by arranging a low-voltage 60-cycle circuit definitely limited to 5 or 10 ma. Make contact on the upper arm with a flexible conducting arm band, and with fingers relaxed and extended, have an assistant explore the forearm with a small blunt electrode connected to the other side of the circuit. Locations will be found where the individual fingers will be contracted, very much to the surprise of the subject.

The muscular contractions produced by currents in excess of one's let-go current can be very powerful. In 1942, a shipyard workman in the San Francisco Bay area attempted to obtain a better view of an in-plant Christmas Eve party by climbing around an overhead crane. He slipped and, to keep from falling, grasped the 440-volt crane trolley wires. He immediately froze to the trolley wires, and his body swung back and forth until someone opened the switch. He then fell to the ground, fractured his skull, and died. A similar accident occurred in 1947 at a brass works in the Los Angeles area. A laborer lost his footing when climbing down from the roof of the factory. As he fell he grasped the 220-volt trolley wires of an overhead crane, and froze to the trolley wires. His body dangled from the wires until one employee opened the switch while another employee caught him as he fell. He received severe burns but recovered.

The foregoing may serve to illustrate that electrical stimulation of the muscles does not promote "grabbing" of contact, but that once a firm grasp of an energized conductor has been established, the resulting muscular contractions may prevent a victim from releasing his grasp of the energized conductor. To prevent the possibility of freezing to a circuit, it is good practice first to touch with the back of the middle finger a supposedly de-energized conductor or the frame of an electrical device.

In reply to Mr. Krasin, the muscular reactions produced by electric stimulation are due to the current actually flowing through the body, and are independent of the voltage per se. For the same circuit and contact conditions, the higher the voltage, the greater the current, and the more violent the muscular reactions. Human tissue has a negative resistance characteristic, i.e., the body resistance decreases with both increasing current and with increasing time of contact, with the result that doubling the voltage more than doubles the current.

On the lower voltages the circuit is frequently not completed until the palm of the hand actually makes contact with the energized conductor or with a defective tool held in the bare hand. The resulting current may effectively freeze the victim to the circuit. Barring fortuitous circumstances, the resulting shock may be of long duration with disastrous results. In contrast, on higher voltages, contact with the circuit may be by a jump spark or arc just prior to obtaining a grasp of the conductor. The resulting violent involuntary muscular reactions may throw the person away from the circuit, thereby quickly interrupting the current and saving the victim. Also, the current flowing through the body in high-voltage accidents is sometimes sufficient to blow fuses or trip circuit breakers, thus decreasing the duration of contact and thereby increasing a victim's chance of survival.

Interruption Tests on High-Voltage Air-Break Contactor

C. A. LISTER
MEMBER AIEE

HIGH-VOLTAGE air-break contactors are frequently used in controllers, both for starting motors and for interrupting short circuits. According to the National Electrical Manufacturers Association (NEMA) standards,[1] 2,300- to 4,600-volt contactors for this duty may be used in circuit locations where the available symmetrical short-circuit kilovolt-amperes (kva) are as high as 50,000 kva. In this paper new contactors designed for this service are described, and the tests made to prove the interrupting rating are discussed.

Paper 56-31, recommended by the AIEE Industrial Control Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Winter General Meeting, New York, N. Y., January 30–February 3, 1956. Manuscript submitted July 11, 1955; made available for printing November 15, 1955.

C. A. LISTER is with the Electric Controller & Manufacturing Company, Cleveland, Ohio.

Contactor Design

The new contactors are rated 200 and 400 amperes continuous, 2,300 to 4,600 volts, 3 phase, 60 cycles, and have an interrupting rating of 50,000 kva available symmetrical. A contactor may be furnished with an a-c or a d-c operating magnet.

The high-voltage parts are enclosed by a phase barrier assembly with closed front and top but open at the rear to vent gases from the arc chutes. This barrier assembly may be drawn out when inspection of the contactor is required. Fig. 1 shows a complete 200-ampere contactor. For Fig. 2 the barrier assembly was removed, and for Fig. 3 the arc chutes were removed.

As shown in the figures, the arc chutes can be removed by releasing the clamps and pulling the chutes out. This permits complete inspection and easy access for replacement of contact tips, a highly desirable feature since contactors are normally serviced in place. As the simplicity of the moving parts suggests, the contactor was designed for a life of a few million operations with only minor maintenance.

Fig. 4 is a side view of a pole assembly with many parts removed to show the current paths. When the contactor is closed, the current passes through a permanently connected blowout coil to the contacts and thence through a flexible connector to the outgoing terminal via an arrangement of conductors designed to reduce the electromagnetic force, which otherwise would cause premature separation of the contacts under fault conditions. When the contactor opens, the arc drawn between the contacts of each pole is forced onto and up arc runners in the arc chute by a transverse magnetic field set up between the blowout ears by the blowout coil. Cooling plates are arranged to force the upward moving arc into a sinuous path, and the lengthening and cooling of the arc results in interruption within the arc chute. Fig. 5 shows a stack of cooling plates and a few separate plates. The design of conductors surrounding the flexible connector was decided upon after a series of calculations and tests in which the arrangement and

orientation of the conductors were varied. To calculate approximately the forces on the connector for a particular arrangement of conductors, the connector and other conductors were replaced by a number of straight-line segments and current was assumed to be concentrated in filaments of very small diameter along these segments. Assuming the flux paths to be entirely in air, the flux density was computed at each of several points along the connector. This was done by using Ampere's law to find the flux density contribution from each of the segments and then superposing the components of flux density. Forces on the connector segments were then calculated by the well-known expression "Force equals flux density times current times length" ($F=BIL$) with the use of the component of total flux density perpendicular to the plane of the connector. The equivalent force tending to separate the tips was then computed.

The calculations showed that with a conductor arrangement of the type shown in Fig. 4 the electromagnetic force tending to separate the tips was a small fraction of the force which would be present for a simple loop arrangement of conductors. To check the calculations, tests were made at high currents and low voltages to find the contact spring force necessary to prevent tip separation. The tests confirmed the calculations.

The single, permanently connected blowout coil was chosen for simplicity and reliability. A 2-coil common-core blowout was considered and tested, but difficulties were encountered at high currents. The 2-coil blowout consisted of a permanently connected section to effect transfer of the arc to the arc runners and an intermittent section of many more turns located alongside the permanent coil and connected between the stationary contact tip and its associated arc runner. After an arc is transferred to the arc runner, the two coils are in series. With the 2-coil arrangement an arc rising from the stationary contact tip and striking the associated arc runner short circuits the intermittent blowout coil and, at sufficiently high currents, the voltage induced in the intermittent coil by the magnetically coupled permanent coil is enough to maintain the arc between the stationary contact tip and its arc runner. Under these conditions the coils are in opposition and the flux pattern is greatly distorted; the arcs fail to move up the runners and the contactor fails to clear the circuit.

The use of a single-coil blowout gives less magnetomotive force in a blowout structure, after arc transfer, than could be obtained with a double-coil blowout. With the same blowout iron, this would give lower flux densities and lower arc velocities resulting in higher arc times, which might be excessive at low currents. However, by using a completely laminated iron structure with parts oriented to provide flux paths predominantly parallel to the laminations in regions of high flux density, and by placing the cooling plates close to the contact tips, reasonably short low-current arcing times have been obtained. For example, the contactor with a 200-ampere blowout coil has an arcing time of about 6.5 cycles for a low-power-factor static load of 40 amperes at 2,300 volts switched by a single pole.

Another factor considered in design was impulse voltage strength. The NEMA standard for future design[1] gives impulse withstand voltages of 45 kv for 2,300-volt controllers and 60 kv for 4,600-volt controllers, the impulses being applied between each circuit and all other circuits and grounded parts, except that the open tip gap of the contacts is excluded. The contactor described has been impulse-tested, and withstand voltages are greater than 60 kv both phase to phase and phase to ground. A large contact opening of $1^1/_4$ inches insures high impulse strength of about 40 kv between incoming and outgoing terminals of the contactor. When impulses are applied between two incoming terminals with the corresponding outgoing terminals tied, so that two gaps are in series, the impulse strength is about 60 kv. To increase further the impulse strength between contacts would require a still larger contact opening and a larger operating magnet. This would increase the closing time of the contactor more than permissible in many applications.

To assure fast opening of the contactor, particularly desirable in applications where emergency stopping of motors is required, and to reduce the effects of the electromagnetic forces during short circuit, the contact spring pressures of the contactor have been made much higher than those for low-voltage contactors of the same current rating. The higher spring pressures also make for better contact of the tips which reduce contact heating and deterioration.

Opening at high speed, the armature tends to rebound from the stop, and the energy-absorbing mechanism shown in Fig. 6 is used to eliminate such rebounds.

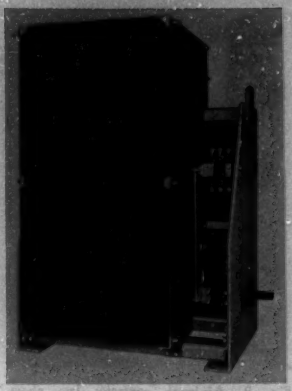


Fig 1. Air-break contactor, 200 amperes, 2,300 to 4,600 volts, 50,000 kva interrupting rating

As the armature first strikes the stop block, the block rotates and rubs against the armature. By the time the armature reaches the flat in the block, much of the energy is absorbed and no perceptible rebound occurs.

Standards for Interrupting Capacity

Since high-voltage contactors are normally used as part of controllers in which current transformers, overload relays, and other accessory equipment are provided, interrupting ratings are given for controllers rather than for contactors alone. NEMA standards[1] cover the appropriate type of controllers; these standards include interrupting ratings and the test procedure used in proving a controller. The same test procedure is commonly used for testing separate contactors as well.

The interrupting rating expresses the maximum available symmetrical fault kva to be permitted at the line terminals of a controller when the controller is expected to clear any fault beyond its load terminals. The controller must be able to interrupt the circuit regardless of the magnitude of any d-c components which may also be present.

In test, a controller is connected in series with reactors and a test generator. The reactors are adjusted so that when the controller is short-circuited at its line terminals, the fault current will correspond to the kva rating of the controller. The circuit is actually closed and voltages and currents are recorded by oscillograph. With the circuit thus

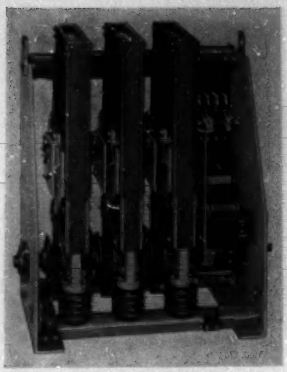


Fig. 2. Air-break contactor less phase barrier assembly

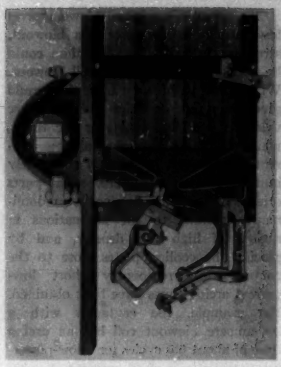


Fig. 4. Cutaway side view of one pole of air-break contactor showing blowout coil contacts, arc runners, and antiblowoff assembly

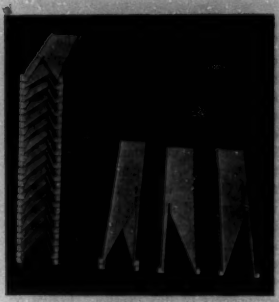


Fig. 5. Cooling plates for arc chute

calibrated, the short circuit is removed from the controller line terminals and is placed at the load terminals, the circuit remaining unchanged otherwise. Three close-open operations at 2-minute intervals are then made in which the contactor of the controller makes and breaks the circuit. Currents and voltages are recorded as in the calibration test.

At the end of the tests, the controller must be able to withstand rated voltage when the contactor is open and to carry rated current at rated voltage for limited time but not necessarily without exceeding rated temperature rise. During tests any arcing to the controller cabinet or other grounded parts constitutes failure to meet the NEMA standards. To detect any such arcing, a fuse is connected between the cabinet and ground and either the fault or the neutral of the power supply, but not both, is grounded.

Test Station

A series of interruption tests were made on the contactor at the short-circuit testing station of Crompton Parkinson Limited in Chelmsford, England, one of the stations of the Association of Short-Circuit Testing Authorities inspected and approved by the British National Physical Laboratory. This station has a short-circuit capacity of 400,000 kva and is well equipped for contactor testing. Switches are provided for connecting the current adjusting reactors, and, for tests at the same voltage, changes in test current value between tests can be made in less than 2 minutes. The oscillograph at this station has 12 channels which permit recording all important variables during contactor short-circuit tests; facilities are provided for developing oscillograph films between shots.

An unusual feature of the station is means for control of recovery voltage. A fault can be initiated during build-up or decay of generator excitation so as to provide a recovery voltage greater or less than the voltage prior to fault, as desired.

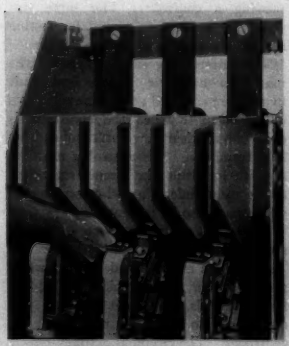
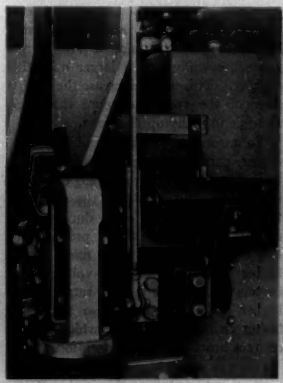
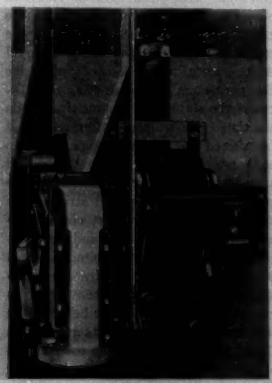


Fig. 3. Partial view of air-break contactor with arc chutes removed



A—Contactor closed



B—Contactor open

Fig. 6. Contactor antirebound mechanism

Table I. Interrupting Tests on 2,300 to 4,600-Volt Controller With Air-Break Contactor

Available* Kva	Operation†	Applied Voltage	Power‡ Factor, Per Cent	Making‖ Current, Peak Amperes	Symmetrical‖ Current, Rms Amperes	Recovery¶ Voltage, Per Cent	Arc Duration, Cycles	Total Break Time, Cycles
35,000	Calibration	2,400	3.5	34,000	8,800			
35,000	CO	2,400	3.5	15,000	7,500	103	0.8	2.3
50,000	Calibration	2,400	4.4	38,000	13,600			
50,000	CO–2 min—CO–2 min—CO	2,400	4.4	19,000	10,300	104	0.7	2.3
60,000	Calibration	2,400	5.0	44,000	15,400			
60,000	CO–2 min—CO–2 min—CO	2,400	5.0	21,000	11,900	103	0.6	2.3
50,000	Calibration	4,700	15.0	14,000	6,600			
50,000	CO–2 min—CO–2 min—CO	4,600	15.0	14,000	6,100	100	0.9	2.6
60,000	Calibration	5,100	15.0	16,000	7,100			
60,000	CO–2 min—CO–2 min—CO	5,000	15.0	16,000	6,700	100	1.0	2.7

* Kva available at controller line terminals.

† CO = close-open, CO–2 min—CO–2 min—CO = close-open–2 minutes—close-open–2 minutes—close-open.

‡ Power factor of circuit ahead of controller.

‖ Maximum of 3 phases.

‖ Average of 3 phases.

¶ Normal frequency recovery voltage in per cent of rated voltage, i.e., per cent of 2,300, 4,600, or 5,000 volts.

Test Results

Table I summarizes the more important data from the test results, and Fig. 7 is a typical oscillogram. The tests recorded are for a complete motor starter including disconnecting switches, current transformers, and overload relays, all enclosed in an indoor-type steel cabinet. Control power was provided by the control transformer of the starter, and tripping of the contactor was caused by operation of the overload relays.

For all tests the controller cabinet was connected to ground by a fine wire which served as a fuse, the cabinet being otherwise insulated from ground. The Y-point of the fault was grounded and the neutral of the generator was isolated. There were no failures to ground occuring.

All tests were performed at the station frequency of 50 cycles. Such tests are considered slightly more severe than tests at 60 cycles.

As the table shows, test duty cycles of three close-open operations at 2-minute intervals were performed at 50,000 kva, 2,300 volts and 4,600 volts and at 60,000 kva, 2,300 volts and 5,000 volts. Since the controller is rated for use where the available fault kva does not exceed 50,000 kva, it was felt that there was little point in making tests much above this value and, therefore, no tests were performed at more than 60,000 kva available. Arcing time for tests at 35,000 to 60,000 kva available was less than 1 cycle and total break time was about 2.5 cycles.

At the end of the tests the controller was in substantially the same condition as at the start and was satisfactory for use without replacement of parts. Certificates have been issued by the testing station which include the pertinent test data and certify that the controller successfully met the interrupting tests specified by NEMA standards.[1]

In addition to the tests of Table I, other tests were made in which the contactor alone was used and separate control power was furnished. Since the disconnects, current transformers, and controller wiring were eliminated, the actual fault currents approached the available fault currents, shown by the calibration tests, more closely than did currents in tests in the complete controller, and the d-c components of fault current decayed less rapidly. For the same available fault kva, the reduction in fault current caused by the controller impedance is, of course, much less in 4,600-volt tests than in 2,300-volt tests.

The normal frequency recovery voltage was no less than rated voltage. As mentioned earlier, this was accomplished by closing the circuit during build-up of generator excitation and by using sufficient steady-state excitation to give more than rated voltage.

In the tests within the interrupting rating of the controller, no premature separations of the contacts occurred. In the tests at 60,000 kva, 2,300 volts, two close-open operations were made without premature contact separation and in the third close-open operation a very brief separation occurred on one pole. This had no adverse affect on the operation of the contactor. No premature separations occurred at 60,000 kva, 5,000 volts. Thus, the contactor parts designed to minimize the electromagnetic forces in the flexible connector performed their function well. As the armature travel record of Fig. 7 indicates, the antirebound assembly was entirely satisfactory during interruption tests.

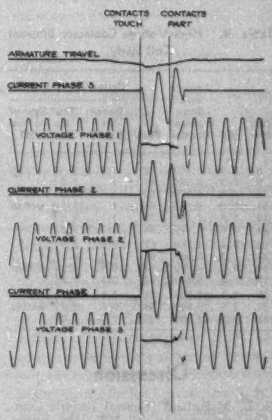


Fig. 7. Oscillogram of interrupting test, 60,000 kva available, 5,000 volts

Conclusions

A new air-break contactor has been developed with high interrupting rating and high impulse strength. This contactor has been designed to have long life under severe operating conditions, and particular care has been taken to make inspection and maintenance easy. Tests have shown that this contactor is suitable for interrupting short circuits in locations where the available symmetrical fault kva is as high as 50,000.

Reference

1. Standards for Industrial Control. Publication no. *ICI-1954*, National Electrical Manufacturers Association, New York, N. Y., May 1954, pt. 24, pp. 1–8.

(*Discussion follows.*)

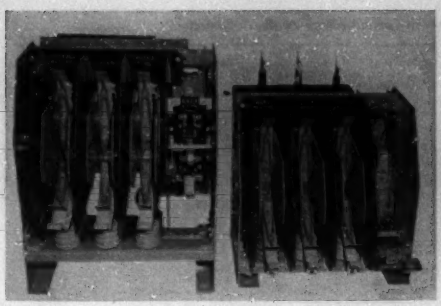


Fig. 8. Front view of old design (left) and new design (right) showing reduction in size. Arc chutes removed

Fig. 9. Proposed "limitamp" panel on test car at Philadelphia short-circuit laboratory

Table II. High-Voltage Contactor Blowout Coil Study

Test No.	Kv	Average Rms Symmetrical Current	Mva	Cycles, 60-Cycle Basis: Transfer Time	Total Arc Time	Type of Blowout Coil
8	5.5	6,500	62	1	2	
9	5.5	6,400	61	1	1.5	
10	5.5	6,500	62	1	2.5	2-section
11	5.6	6,670	65	1	1.5	
12	5.6	6,550	63	1	2.6	
13	5.6	6,670	64	1	1.3	
14	5.4	6,770	63	0.8	1.5	
15	5.6	6,600	64	0.6	0.9	1-section
16	5.6	6,600	64	0.8	1.2	
17	5.5	6,700	64	0.5	1.2	
25	2.7	12,300	58	0.5	0.8	1-section
26	2.6	12,300	56	0.5	0.9	
27	2.6	12,200	55	0.5	1.0	

Discussion

G. L. McFarland (General Electric Company, Schenectady, N. Y.): The author has commented on the 2-coil and single-coil blowout construction, and has stated that the contactor fails to clear the circuit when a 2-section coil blowout is used.

In our experience there has been no evidence that the double-section blowout coil is incapable of handling the same kva rating as the single-section design. Our tests have not substantiated the effect on the ability to interrupt as the author suggests.

Since about 1950, the General Electric Company has been manufacturing an air-break contactor somewhat similar to the one the author describes. Goldberg reported this in 1951.[1]

Recently, a modification of this design was introduced which is approximately 25 per cent smaller than the previous unit. It has dead-back construction, and it is front-connected. Fig. 8 shows the old and new designs.

Fig. 9 shows a high-voltage starter on a test car being moved to a test cell. For these tests single-section blowout coils were compared with 2-section blowout coils which have been in use for several years.

Table II shows the effect of a single-section blowout coil on interruption. The time taken for the arc to transfer to the arc runners as well as the total arc time was decreased by the use of the single-section blowout coil. Since the performance of the contactor was improved, a single-section blowout coil will be used in the future on this device. Table II shows data for both types of blowout construction taken at 60 mva.

It should be noted that these tests have been made in accordance with the revised NEMA standard for high-voltage high-interrupting-capacity controllers. In this standard, the circuit is set up with the controller short-circuited at its line terminals in such a way that its d-c component decay time is not less than that indicated in Table III.

Table III. D-C Component Decay Time

Time After Initiation of Short Circuit, Cycles, 60-Cycle-per-Second Basis	Ratio of Total Rms Current in Phase with Maximum D-C Component to Rms Symmetrical Current Corresponding to Interrupting Rating, Kva
1/2	1.6
1	1.4
2	1.2
3	1.1
4	1.0

Reference

1. A New 5,000-Volt Air-Break Contactor for Industrial Service, Leon J. Goldberg. *AIEE Transactions*, vol. 70, pt. II, 1951, pp. 2030–34.

C. A. Lister: It is interesting that Mr. McFarland has found improved performance when a single-section blowout coil is used and that the double-coil design is being abandoned.

The tests listed by him on 2-coil blowouts do not include any at currents of over 6,670 amperes. The failures referred to in the paper were for slightly greater currents and 2,300 volts, corresponding to about 30,000 kva. No difficulties were experienced in tests at 2,500 amperes and no tests were made at intermediate currents. Thus, the lowest current value at which arcs fail to move up the arc runners was not found with any precision. The longer arcing times reported by Mr. McFarland for the 2-coil blowouts, in spite of presumably greater blowout flux density, appear to indicate approach to the sort of difficulty mentioned in the paper.

Table III will appear in revised NEMA Industrial Control Standards, Part 24, to be published in 1956. While the contactor and controller discussed in the paper were tested long before this table was proposed, the tests were severe enough to fulfill essentially the requirements of the revised standards.

Economic Studies Applied to Industrial Power Systems

H. B. BACKENSTOSS
MEMBER AIEE

INDUSTRIAL plants are quite understandably governed principally by the problems associated most directly with production. It is important, naturally, to concentrate the talents of plant management on the larger elements in the cost of the finished product because new ideas can be very effective here in creating savings in cost or increasing output. It is therefore not surprising that the cost of electrical service, particularly in respect to investment requirements for the necessary physical plant, is likely to receive less attention. A procedure often adopted in plants where power does not enter heavily into the cost of the finished product is to plan new installations or make changes in production facilities with only incidental regard to the problems of electrical supply and thereby oblige the electrical planning engineer to furnish power to the production equipment as best he can. This does not always lead to an electric system of optimum cost. It may not lead to minimum long-term cost of the total production facility.

There is a second factor. Even when power system costs are recognized in production planning, immediate costs are apt to receive the greatest attention. If future costs are not thoroughly studied and given their due weight, an electric system of low first cost but higher future cost is in danger of selection. The advantages of an alternative system of greater initial cost but favorably lower future costs due, let us say, to wise provision of suitable margin for growth, may go unrecognized. Of course there is the desire of management to use its dollars in production rather than in service equipment and when the supply of dollars is limited, as it almost surely is, this element must be recognized in power system planning. A greater obstacle to securing optimum cost sometimes lies in the difficulty of evaluating future costs. With an adequate method at hand to accomplish this, an industrial power system could be planned with reasonable assurance of securing minimum long-range cost.

A method of evaluating future costs and giving them their proper weight against present costs is presented and explained in the following, together with several examples to illustrate its application.

Paper 56-22, recommended by the AIEE Industrial Power Systems Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Winter General Meeting, New York, N. Y., January 30-February 3, 1956. Manuscript submitted August 8, 1955; made available for printing November 7, 1955.

H. B. BACKENSTOSS is with Jackson and Moreland, Boston, Mass.

Nature of Economic Analyses

Perhaps one of the best ways of defining an economic study is to begin by saying what it is not. It utilizes construction cost estimates yet it is more than an evaluation of initial cost. It considers later costs yet is not an analysis which reduces the uncertainties or the intangibles of the future to certainties and measurable quantities of the present. It is not an analysis which can or even pretends to put a dollar sign on such things as ease of system operation or on psychological factors which are often important considerations in the acceptability or ultimate real cost of an electric system. In brief, it is not a device for reducing all the present and future differences between one or more alternate plans of system development to the common denominator of the dollar sign. The system having lowest cost through its useful life, as determined by careful analysis, is not necessarily the system which must be chosen willy-nilly. Judgment, both engineering and management judgment, is still required.

The long-range economic study will, however, isolate all the known elements of cost which would be incurred by the several appropriate power systems and express them on the same basis as closely as possible. Usually the investment on which annual depreciation charges, income and property taxes, and investor's return are incurred, and the annual operating expense, comprised of labor, electrical losses, and maintenance are the items which must be considered. Differences in fixed charges and operating expense for the several development plans will usually continue throughout the useful life of the systems. An economic study of the type described here accumulates these differences and expresses them in terms of today's dollars by application of discount, or present worth, factors. By thus considering all the known elements of cost, by using the best possible cost estimates, and by converting future cost differences to present values, the economic study provides a powerful guide to effective long-range planning. The area of the unevaluated is reduced; sound mechanics in establishing a logical conclusion from a given set of assumptions is substituted for clairvoyance. Judgment is obliged to work only within a greatly narrowed field of uncertainty; the chances of selecting the plan best suited to the circumstances are greatly improved.

Specific Features of Problem

Adequate planning implies features which hardly require mention. An industrial power system should not only carry existing loads without restrictive bottlenecks but also have capacity for future growth. It should have a degree of reliability commensurate with the industrial process it serves. It should be flexible to meet shifts in loads from one area to another. It should be arranged for maintenance without excessive cost or intolerable interruptions to service. Reduced to specific terms, this means proper voltage, proper transformer sizes, adequate rupturing duty in circuit breakers, suitable wire sizes, and effective circuit arrangements for load protection. If the plant generates its own power supply either in whole or in part, there are the additional considerations that unit size and throttle steam conditions must be suitable to both the plant electric load and the plant steam load. All these elements of choice are subject to economic studies because the question which invariably arises in any industry which hopes to expand is whether to spend extra money in the present to secure advantages in capacity or flexibility or reliability against some future need or to wait until the future need actually arises and meet it by a possibly greater expenditure at that time.

It is usually found that additional capacity can be added at a favorably low incremental cost when a facility is built and that it costs much more if added by later modifications. Extra capacity provided in the initial installation is termed "prebuilt" capacity in the following discussion. Capacity added by modifications

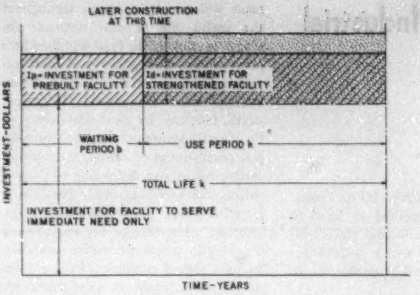

LATER CONSTRUCTION AT THIS TIME
I_p = INVESTMENT FOR PREBUILT FACILITY
I_d = INVESTMENT FOR STRENGTHENED FACILITY
INVESTMENT-DOLLARS
WAITING PERIOD b
USE PERIOD h
TOTAL LIFE k
INVESTMENT FOR FACILITY TO SERVE IMMEDIATE NEED ONLY
TIME-YEARS

Fig. 1. Investment–time relation

when the future need arrives is termed "deferred" capacity. Should it happen that prebuilt capacity costs more than deferred capacity the designer's choice becomes simple. It is then obviously not advantageous to build any extra capacity until the need actually arises; the minimum immediate cost solution is best. This paper is not concerned in cases where prebuilding of capacity is more costly than deferring it.

Development of Equation

The basic mathematical statement of this problem can readily be developed from a familiar operation, that of selecting the economic size of transformer for a substation which is expected eventually to serve a load greater than its immediate load.

Assume that a step-down substation is required to serve immediate load L_1. The designer knows (or predicts) that L_1 will step up to a larger load L_2 after b years. A transformer able to handle L_2 can be installed now with no later changes required or a transformer to serve only L_1 can be selected, with future strengthening, say, by a parallel transformer, at the end of b years to handle L_2.

It should be noted that in practice a load increase would almost surely be achieved gradually. Representing it by a 1-step increase greatly facilitates understanding of the mathematics of the problem and does not impair the applicability of the equations to the actual case of gradual increase.

Fig. 1 shows the investment–time relation corresponding to this problem. The transformer able to serve only L_1 is represented by "investment for facility to serve immediate need only." The cross-hatched area I_p corresponds to the excess investment required for the transformer needed to serve L_2 rather than L_1. I_p continues for the total life of the equipment, k years, but is "unused" during the waiting period of b years before L_2 appears. It is fully loaded thereafter during the h year use period. The dotted area I_d corresponds to the investment to modify facilities originally built to serve L_1 but which after b years are required to serve L_2.

Case A: Facility Repeatedly Renewed at Original Cost

Case A assumes that each investment has the same total life k, commencing on the date the investment is made, and that there is no change in price level with time. Each facility or facility addition can be renewed independently as it reaches the end of its own life and is renewed indefinitely into the future. Investment cost thus remains level after load L_2 has developed.

The annual cost increase incurred by prebuilding to handle L_2 is

$$C_p = \rho I_p + \Delta O_p \tag{1}$$

Similarly the annual cost increase incurred when facilities initially built only to serve L_1 are strengthened to handle L_2 is

$$C_d = \rho I_d + \Delta O_d \tag{2}$$

In these equations

C = total annual costs of capacity installed to serve $(L_2 - L_1)$
ρ = fixed charge rate on investment
I = investment for the capacity installed to serve $(L_2 - L_1)$
ΔO = differential operating charges (labor, electrical losses, etc.)
p = prebuilt
d = deferred

Depreciation charges are included in ρ. Since each investment in case A is assumed to have the same total life k, depreciation and ρ have an identical value for all.

The differential operating charge terms appear in equations 1 and 2 in order to recognize the savings which sometimes have an effect on the economic standing of alternate plans. It is impractical and unnecessary to deal with total operating costs, hence ΔO_p is only the difference in cost effected by serving L_1 from a facility designed for L_2. Thus ΔO_p lasts only through the waiting period b and is zero thereafter. Similarly ΔO_d is zero until the waiting period ends and has a value equal to the difference in cost of the strengthened facility compared with the prebuilt facility in serving L_2 thereafter. Often this difference is negligible.

By virtue of the assumption of repeated renewal of facilities, the investment terms in equations 1 and 2 continue as perpetual costs. Investment I_p commences at the date of initial construction and I_d commences after the waiting period b.

The total of excess annual costs for the alternate plans may be brought to a common base for comparison by applying annuity and present-worth factors to convert them into dollars as of the date of initial construction.

Present worth of total excess cost of prebuilding

$$= \sum_{0}^{\infty} P_{1,s,n} C_p = A_\infty \rho I_p + A_b \Delta O_p \tag{3}$$

Present worth of total excess cost of deferring

$$= \sum_{0}^{\infty} P_{b,b+1,n} C_d = P_b A_\infty (\rho I_d + \Delta O_d) \tag{4}$$

where

$P_{1,s,n}$ = present worth of \$1 payable at the end of 1,2,n years with interest at rate r
$A_{b,h,k}$ = present worth of an annuity of \$1 per year for b, h, k years with interest at rate r. A_b and A_k appear later
A_∞ = present worth of a perpetual annuity of \$1 per year with interest at rate r
r = component for interest or return on money included in the total fixed charge rate

A perpetual annuity A_∞ does not have infinite value but is equal to $1/r$. Thus if r equals 6 per cent (%), A_∞ equals \$1/0.06, or \$16.67. Substituting this in equations 3 and 4 gives simpler expressions

$$\sum_{0}^{\infty} P_{1,s,n} C_p = \frac{\rho I_p}{r} + A_b \Delta O_p \tag{5}$$

$$\sum_{0}^{\infty} P_{b,b+1,n} C_d = \frac{P_b}{r} (\rho I_d + \Delta O_d) \tag{6}$$

These equations show the economic criterion for prebuilding. If the present worth of the accumulated excess annual costs for prebuilding given by equation 5

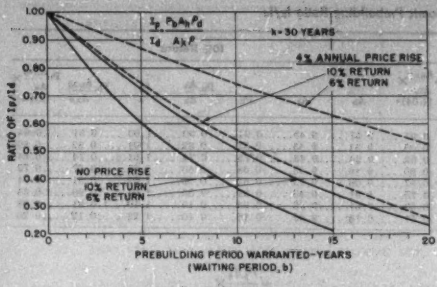


Fig. 2. Period for which prebuilding is warranted, case B

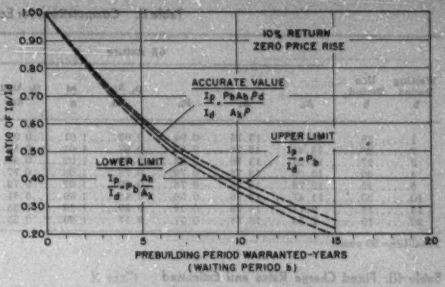


Fig. 3. Lower and upper limits for the prebuilding time warranted

exceeds the present worth of the accumulated excess annual costs for deferring given by equation 6, initial construction should be limited to immediate needs. If the reverse is true and the costs for deferring are greater, initial construction should provide for future needs as well.

The relationship between equations 5 and 6 may be expressed as a ratio. When prebuilding and deferring are equal in cost

$$\frac{\frac{\rho I_p}{r}+A_b\Delta O_p}{\frac{P_b}{r}(\rho I_d+\Delta O_d)}=1 \tag{7}$$

or

$$\frac{I_p}{I_d}=\frac{\frac{P_b}{r}\left(\rho+\frac{\Delta O_d}{I_d}\right)}{\frac{\rho}{r}+A_b\frac{\Delta O_p}{I_p}} \tag{8}$$

Case B: Facility not Renewed or Renewable at End of Life

In many situations an investment to strengthen facilities has an abbreviated life and must be depreciated at an accelerated rate in order to retire it with the facility of which it is a part. Moreover the cost of replacing facilities when they die may be higher or lower than the cost of the facilities originally, due to changes in price level. These two considerations lead to terminating the economic comparison at the end of the life of the original facilities. Increased depreciation charges will be incurred on the deferred investment commensurate with its shortened useful life.

Table I. Price Rises

	Index, January 1935	Index, January 1955	Ratio	Compound Rate of Increase
Overhead conductors and devices	106	420	2.53	1.048
Power transformers	200	361	1.80	1.030
Oil switches	221	423	1.91	1.033
No. 6 weatherproof copper wire	122	288	2.36	1.044
Turbogenerator units	289	640	2.21	1.040
Electrical labor	369	606	2.48	1.046

The total fixed charge factor ρ includes four principal elements

$$\rho=(e+i+r+d) \tag{9}$$

where

e = expense rate for local taxes, insurance, that portion of maintenance which is proportional to investment, etc.
i = required return on investment for income taxes (federal, state)
r = rate of return on invested funds (interest, dividends, etc.)
d = depreciation rate, sinking fund basis

Total fixed charges on deferred investment will be

$$\rho_d=(e+i+r+d_h) \tag{10}$$

which will allow for retirement at the end of h years. Fixed charges on the prebuilt investment, on the other hand, will be

$$\rho_p=(e+i+r+d_k)=\rho \tag{11}$$

which recognizes the normal life of k years for prebuilt facilities. Equation 11 gives the value applicable for fixed charges throughout case A, in which the term ρ without subscript was used. For continuity, ρ will continue to be used here to apply to facilities having normal life k.

The effect of an allowable depreciation rate for income tax purposes which exceeds the sinking fund depreciation rate appearing in equations 9, 10, and 11 is worth digressing a moment to examine. By whatever method depreciation is handled, the total at the end of the equipment life is, of course, always the same sum, namely the initial investment. The use of a depreciation rate for tax purposes which exceeds true economic depreciation reduces the taxes a company must pay the government during the early years of an investment's life. Because taxes saved in this fashion are repaid in the later years of the equipment's life, this amounts to receiving a temporary loan from the government at zero interest. Its effect therefore is to make the true rate of interest (or dividends) on borrowed funds less than the apparent rate. Recent statutory changes in federal income tax laws, providing for sum-of-the-digits and declining balance methods of computing allowable depreciation, have accentuated this effect.

The annual cost increase incurred in case B by prebuilding to serve L_2 is the same as before

$$C_p=\rho I_p+\Delta O_p \tag{12}$$

The annual cost increase incurred when deferred facilities are constructed to serve L_2 differs from case A in its requirement for increased depreciation, and is

$$C_d=\rho_d I_d+\Delta O_d \tag{13}$$

Bringing these excess costs to a common base for comparison by applying appropriate annuity and present worth factors gives

Present worth of total excess cost of prebuilding

$$=\sum_0^b P_{1,2,k}C_p=A_{2\rho}I_p+A_b\Delta O_p \tag{14}$$

and

Present worth of total excess cost of deferring

$$=\sum_0^b P_{b,b+1,k}C_d=P_bA_b(\rho_d I_d+\Delta O_d) \tag{15}$$

As before, with equations 5 and 6 in case A, the relation of equations 14 to 15 determines the economic choice between prebuilding and deferring

The ratio at which prebuilding and deferring are equal in cost is

Table II. Computations for Economic Prebuilding Ratio I_p/I_d

Waiting Period b	Use Period h	6% Return A_b	6% Return A_k	6% Return P_b	6% Return $P_b \frac{A_b}{A_k}$	6% Return $\frac{\rho_d}{\rho}$	6% Return $P_b \frac{A_b \rho_d}{A_k \rho}$	6% Return $P_b \frac{A_b \rho_d}{A_k \rho} \times (1.04)^a$	10% Return A_b	10% Return A_k	10% Return P_b	10% Return $P_b \frac{A_b}{A_k}$	10% Return $\frac{\rho_d}{\rho}$	10% Return $P_b \frac{A_b \rho_d}{A_k \rho}$	10% Return $P_b \frac{A_b \rho_d}{A_k \rho} \times (1.04)^b$
1	29	13.59	13.76	0.94	0.93	1.01	0.94	0.97	9.37	9.43	0.91	0.90	1.00	0.91	0.94
2	28	13.41	13.76	0.89	0.87	1.01	0.88	0.95	9.31	9.43	0.83	0.82	1.01	0.82	0.89
3	27	13.21	13.76	0.84	0.81	1.02	0.82	0.92	9.24	9.43	0.75	0.74	1.01	0.74	0.84
4	26	13.00	13.76	0.79	0.75	1.03	0.77	0.90	9.16	9.43	0.68	0.66	1.01	0.67	0.79
5	25	12.78	13.76	0.75	0.69	1.03	0.72	0.87	9.08	9.43	0.62	0.60	1.02	0.61	0.74
10	20	11.47	13.76	0.56	0.46	1.09	0.51	0.75	8.51	9.43	0.39	0.35	1.04	0.36	0.54
15	15	9.71	13.76	0.42	0.29	1.19	0.35	0.63	7.61	9.43	0.24	0.19	1.10	0.21	0.38
20	10	7.36	13.76	0.31	0.17	1.39	0.23	0.51	6.14	9.43	0.15	0.10	1.22	0.12	0.26

Total life = 30 years.

Table III. Fixed Charge Rates and Estimated Values

		10% Return	6% Return
Expense	local taxes	2.0	2.0
	insurance	0.75	0.75
	maintenance	1.0	1.0
Required return for income tax	federal	10.83	4.6
	state	1.08	0.46
Rate of return on funds		10.0	6.0*
Depreciation, sinking fund basis, 30 years		0.61	1.27
Total for ρ		26.27	16.08

* Including 50% bonds at 3.5% interest.

$$\frac{A_k \rho I_p + A_b \Delta O_p}{P_b A_b (\rho_d I_d + \Delta O_d)} = 1 \tag{16}$$

or

$$\frac{I_p}{I_d} = \frac{P_b A_b \rho_d + P_b A_b \frac{\Delta O_d}{I_d}}{A_k \rho + A_b \frac{\Delta O_p}{I_p}} \tag{17}$$

Equations 7 and 8 apply, then, where each investment has a service life independent of associated facilities, and where repeated renewal is necessary to maintain the service and is possible at original cost. Equations 16 and 17 apply, on the other hand, where both the prebuilt and deferred facility reach the end of their life at the same time. From then on the differences between the two which made one plan preferred over the other are erased. The planner regains his freedom of choice.

Special Cases

Case Where ΔO_d = Zero

In many instances the prebuilt facility has little or no advantages over the strengthened facility in operating cost to carry load L_2. The term ΔO_d is therefore zero, or so nearly so it may be neglected. Only the operating differential during the waiting period b exists (and this may be in a direction to assist prebuilding). Equations 8 and 17 for the break-even point then become the following, respectively.

Case A

$$\frac{I_p}{I_d} = \frac{\frac{P_b}{r}\rho}{\frac{\rho}{r} + A_b \frac{\Delta O_p}{I_p}} \tag{18}$$

Case B

$$\frac{I_p}{I_d} = \frac{P_b A_b \rho_d}{A_k \rho + A_b \frac{\Delta O_p}{I_p}} \tag{19}$$

Case Where ΔO_p and ΔO_d = Zero

Another special case is represented by facilities for which operating cost is essentially zero and investment differentials alone control the selection. Circuit breakers are an instance in point. The term ΔO_p becomes zero as well as the term ΔO_d and equations 8 and 17 become the following, respectively.

Case A

$$\frac{I_p}{I_d} = \frac{\frac{P_b}{r}\rho}{\frac{\rho}{r}} = P_b \tag{20}$$

Case B

$$\frac{I_p}{I_d} = \frac{P_b A_b \rho_d}{A_k \rho} \tag{21}$$

Approximate Value of I_p/I_d at Balance Point

On account of the practical difficulties to repeated renewal of facilities at their original price, as assumed for case *A*, the planner is usually more concerned with the equations which compare costs only to the end of the life of the original facilities, as in case *B*. The limits within which the value of I_p/I_d will fall for this case may be gauged by a recasting of equation 21. Operating differentials are for the moment disregarded.

Expanding the terms ρ and ρ_d gives

$$\frac{I_p}{I_d} = \frac{P_b A_b}{A_k}\left(1 + \frac{\rho_d - \rho}{\rho}\right) \tag{22}$$

$$= \frac{P_b A_b}{A_k} \times \left[1 + \frac{(e+i+r+d_b) - (e+i+r+d_k)}{\rho}\right] \tag{23}$$

$$= \frac{P_b A_b}{A_k}\left(1 + \frac{d_b - d_k}{\rho}\right) \tag{24}$$

Depreciation is often a minor element in total fixed charges and the term $(d_b - d_k)$ is even smaller, but the value of the parenthetical expression must be greater than unity, if only by a small amount. therefore I_p/I_d can never be less than $P_b A_b/A_k$ in value and, in fact, approaches that value as its lower limit

The upper limit of I_p/I_d is determined by the terms A_b and A_k. The period h is always shorter than the period k, hence A_b is always smaller than A_k. However, for any given waiting period b, A_b and A_k approach each other in value as the life k becomes longer. (as k lengthens so does h.) In the limit their ratio becomes unity. The parenthetical expression becomes unity at the same time and consequently I_p/I_d takes the value P_b. Therefore for a given waiting period b at the end of which load L_2 is expected, the value of the ratio I_p/I_d for equal cost will lie between $P_b\ A_b/A_k$ as a lower limit and P_b as an upper limit. This relationship can therefore be used for quick determinations of the economics of prebuilding in cases where accelerated depreciation is a factor. Only the values of r, h and k need be known to fix both limits. A knowledge of ρ_d is needed in addition to fix the precise value of I_p/I_d.

Operating Cost Differences

The preceding analysis omits consideration of operating differentials. They will not be treated in detail here. If they are important in determining the economic solution, they may be handled as "equivalent investments." First, the actual differential, for the period it survives, must be converted to an equivalent uniform annual value covering the same

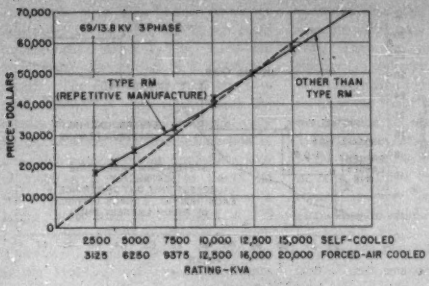


Fig. 4. Power transformer price

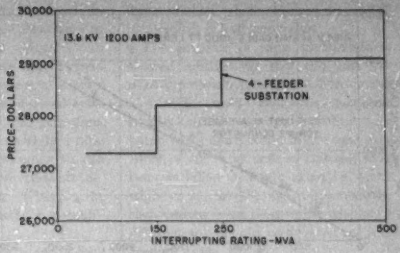


Fig. 5. Metal-clad switchgear price

period as the investment differential (if not already in that form). This may be accomplished by application of annuity factors, $A_{b,\,\rho,\,k}$, for the appropriate periods The equivalent uniform operating differentials so found are next converted to equivalent investments by dividing by the fixed charge rate ρ or ρ_d. Investments so found are, of course, fictitious, but when added to or subtracted from the actual investments as the case may be, a value will result having fixed charges alone equal to the actual fixed charges plus actual operating differential. Equations 20 or 21 can then be used to determine whether prebuilding is warranted.

Multiple Increases in Facilities

In many instances more than one increase in facilities is contemplated. A year-by-year solution in which costs are determined for each year for each proposed development plan and are multiplied by the suitable present-worth factor (P_n) to refer them to the initial construction date is one method of solution. Such a method gives directly the excess of costs of one plan of development versus another. This is often more informative to management than to determine the economic break-even point. The equations developed in the foregoing lend themselves to such a presentation if desired. Care must be exercised to carry such a comparison far enough into the future to reflect all important cost differences and to end it with plans as nearly equivalent in service characteristics as possible.

Price Level

Probably the first questions which come in a planner's mind are: "What effect does a rising price level have on my planning? Is it prudent to prebuild in anticipation of a future price rise? Or is it only a waste of funds during the waiting period never fully recovered despite higher subsequent construction costs?"

The answers to these questions are complex. Many factors enter. One of them, the advance of the art, is quite as important as the rise in general price level, for it determines the real cost of deferred facilities. Suppose that, for the purposes of demonstration, the rise in price level only is considered.

In the 20-year period since 1935, the price rises shown in Table I have occurred in some of the major elements of electric systems. The figures are for public utility construction, but industrial construction should not be far different.

The increases are not all the same, nor have they taken place uniformly with time, but it may be said that since 1935 the annual rise in electric system construction cost has averaged about 4%. Construction deferred for 5 years stands a chance then, if this trend continues, of costing 22% more than its estimated cost today, this being the increase for 5 years taken at 4% compounded annually.

Since an economic balance as determined by equations 8 and 17 is valid only for the actual costs incurred, it is obvious that I_d must be the cost at the price level prevailing when the deferred construction takes place. In making cost estimates, however, I_d is usually measured at the same price level as I_p, the immediate construction.

Let

e = annual rise (long-term trend) in price level

$(I_d)_b$ = cost of deferred construction as estimated at current price level

Then

$$I_d = (1+e)^b (I_d)_b \tag{25}$$

Substituting in equation 21 for the break-even point

$$I_p = (1+e)^b (I_d)_b \frac{P_{b,k,\rho_d}}{A_{k,\rho}} \tag{26}$$

or the prebuilding cost which is warranted economically in a period of rising prices is $(1 + e)^b$ times as great as that warranted in a period of static prices.

The equation for the present-worth factor P is

$$P_n = \frac{1}{(1+r)^n} \tag{27}$$

the inverse in form of the cost trend factor $(1 + e)^n$. If e were equal in magnitude to r the two factors would cancel in equation 26 and the economic prebuilding cost would diminish much more slowly as the waiting period, b lengthened. As it happens, the long-term price rise has in the past been less than the cost of money r so it has not generally been worth while to spend as much to prebuild facilities as equivalent facilities would cost if deferred. Price rise is not, however, a negligible factor.

Illustrative Applications

Fig. 2 illustrates for case B the relationship of I_p to I_d at the economic break-even point as a function of the length of waiting period. The curves have been computed from equation 21 with a total life k of 30 years and various waiting periods b. Values for the curves of Fig. 2 are shown in Table II. Two rates of return on money have been used. The lower, 6%, is typical of public utility experience while the upper, 10%, is more representative of industrial practice. Corresponding total fixed charge rates, including some estimated typical values, have been used in the computations as shown in Table III. Fixed charges on deferred investment ρ_d have been computed for each year with suitable adjustment to allow for accelerated depreciation.

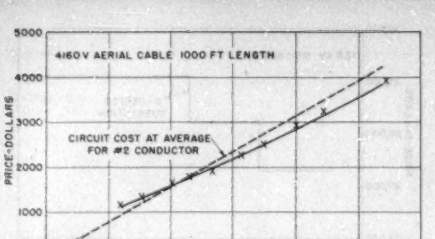

4160 V AERIAL CABLE 1000 FT LENGTH
CIRCUIT COST AT AVERAGE FOR #2 CONDUCTOR
PRICE-DOLLARS
LOAD SERVED AT RATED CURRENT-KVA
WIRE SIZE

Fig. 6. Cost of distribution circuit

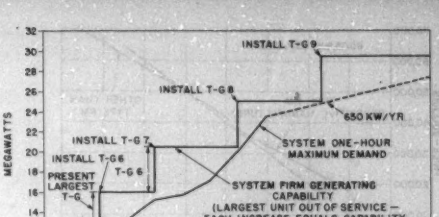


Fig. 7. Optimum program of generator installation

The solid lines of Fig. 2 show the period for which it is economical to prebuild facilities for various ratios of I_p/I_d with constant price level. The dashed lines show the much longer economic prebuilding time with an assumed 4% annual price rise. These curves include investment charges only and omit operating cost differentials. They are presented as typical curves although, of course, they are applicable for the stated rates of return and fixed charges in cases where operating differentials do not exist or where they are of minor importance.

Fig. 3 has been prepared to illustrate the statement that the value of I_p/I_d for equal cost lies between the value $P_b A_h/A_k$ as a lower limit and the value P_b as an upper limit. The solid line is identical with the 10% return, zero price rise curve of Fig. 2. The dashed lines define the upper and lower limits. It will be seen that the margin of variation is not large and that any of the equations would, in this instance, yield a value for economic investment ratio sufficiently accurate for preliminary evaluations. The upper limit P_b is the value applicable to case A as shown by equation 20, which demonstrates the rather small difference between it and case B for the assumed conditions.

Consider the case of an industrial company planning to install its own substation to take power from the local utility. The cost of the power transformer only for a 69/13.8-kv installation consisting of one 3-phase unit is shown in Fig. 4. The immediate load is 12,000 kva and the company can foresee an increase to 18,000 kva in the future. A transformer to meet immediate load would be of 10,000-kva nominal rating costing approximately $39,800. A transformer for the ultimate rating would be a 15,000-kva nominal rated unit costing $57,600. The additional investment for prebuilding to serve the future load would thus be $17,800, i.e., $57,600 − $39,800. Deferring construction until load increased would necessitate installation of a second transformer rated at 5,000 kva nominally and costing $25,250. The reason that this is more costly than the same capacity purchased as an increment to the 10,000 kva unit is well known. It is illustrated graphically in Fig. 4 by the dashed line giving transformer price as a function of size if all sizes were of the same unit price per kva as the 10,000 kva unit. The illustration is slightly less marked than normal because of the step up in price at 10,000 kva in going from repetitive manufacture to nonrepetitive manufacture units.

The ratio I_p/I_d for this illustration is $17,800 divided by $25,250 or 0.71. Fig. 2 shows that it would be economical to prebuild for as much as 3.3 years in advance of the load if money is worth 10% and 5.0 years if it is worth 6%, zero price rise being assumed. Corresponding figures for a 4% annual price rise would be 5.5 years and 9.5 years respectively, but these periods are somewhat long because power transformer prices have actually trended upward only about 3% annually since 1935. In a precise analysis operating differences should be taken into account because electrical losses in the larger transformer during the waiting period unless they should be the same as losses in the smaller unit, would influence the solution. This company would probably find, however, that their load would have to increase within about 5 years or less, assuming money is worth 10% to them, before prebuilding would be warranted.

In selection of interrupting rating for switchgear another instance is found where economic comparisons are helpful. Fig. 5 shows the price of a 4-feeder 13.8-kv metal-clad switchgear unit having interrupting rating ranging from 150 to 500 mva. Using the transformer installations just mentioned it is found that circuit breakers of 150 mva would be adequate for a 12,000-kva substation, whereas 250-mva breakers would be required for an 18,000-kva substation. If prebuilt transformer capacity is not installed, the designer may still have the choice of specifying 250-mva breakers immediately or selecting 150-mva breakers and adding a transformer reactor when load increases. A cost of $3,300 for the necessary reactor has been estimated, with $2,850 for I_d after deducting for two future feeders at 150 instead of 250-mva interrupting rating. Fig. 5 shows I_p equal to $900, i.e., $28,200 minus $27,300, so the ratio I_p/I_d is 0.32. Reference to Fig. 2 shows that the prebuilding period warranted for this ratio is well beyond 12 years under any assumptions as to cost of money or price rise, and that surely it would be economical to provide breakers of a 250-mva interrupting rating in the immediate installation. This example is borne out in practice. In fact, the advantages of excess rupturing capacity in allowing the planning engineer greater freedom of design in the unforseen future are so great that the further expenditure to secure breakers of 500 mva would very likely be made.

Table IV. Relative Cost of Turbogenerator Units

Name Plate Rating, Kw	Assigned Capability, Kw	Purchase Price Index	Price Per Name Plate Kw, Dollars
2,000	2,250	1.00	60.00
2,500	2,813	1.15	54.80
3,000	3,375	1.31	51.70
4,000	4,500	1.61	47.80
5,000	5,625	1.90	45.20
6,000	6,750	2.19	43.50
7,500	8,440	2.64	41.90
10,000	11,250	3.41	40.70

Analysis of a distribution circuit is illustrated in Fig. 6 which shows costs representative of a typical 4,160-volt aerial cable installation of a 1,000-foot length. The curve shows only the cost of the cable. A very rough check indicates that the erected cost would vary in approximately the same proportion as the cable cost. For an immediate load of 1,000 kva, no. 2 cable would be selected at a cost of $1,627. For a larger future load of 1,500 kva, a 2/0 cable costing $2,271 would be required. Therefore I_p equals $644. Should the future load be met by deferred construction it would be necessary to parallel the initial no. 2 cable with another of the same current-carrying capacity in order to secure proportional load division. The deferred cost I_d would be $1,627 and I_p/I_d would be 0.40. Fig. 2 shows that if the load increased at any time prior to 9 years (10% return, zero price rise) it would be worth while to prebuild in anticipation of the increase. Savings in electrical losses during the waiting period would moreover probably noticeably lengthen this 9-year period.

Application to Actual Problem

Actual planning problems solved by the present-worth method are not usually as elementary as the preceding illustration, nor are the results always expressed in the same manner. In the instance reviewed below, the solution was not expressed in terms of ratio I_p/I_d but as the difference in sum of present worths of future costs for the alternative schemes.

A refinery with a present load of 11,500 kw and a 5-unit generating plant of 12,000-kw firm capability was faced with substantial load increases. The largest units in the plant were two 4,000-kw units. The question confronting the planners was what size of generating unit to select for subsequent units.

The most probable load increase for the future was as shown in Fig. 7. It will be noted that the increases occur in amounts which vary widely from year to year. They are related primarily to changes in refinery processing. Prices were determined for turbogenerator units covering the range of sizes deemed sufficiently wide to include the optimum. Table IV gives these prices and their ratio to the price of the smallest unit. It can be seen that the cost per kilowatt decreases as size increases. This is the characteristic which encourages prebuilding.

A schedule of turbine additions for each size of unit was developed against the load curve. The schedules allowed for outage of the largest unit. Plans utilizing machines larger than the existing 4,000-kw units were therefore unavoidably penalized by the requirement of carrying greater investment in spare capacity. Fig. 7 shows the installation schedule for 4,000-kw units.

The present worth of the annual costs for each unit in each schedule throughout its expected life was next determined. These were then totalled for the several schedules by use of present-worth factors selected for the years in which the individual units were installed. The schedules were carried to the point where each had approximately the same installed firm capacity. This is important because great differences in cost can often creep into these comparisons through slight differences in installed growth potential at the end of the period examined in detail. In this problem, return on money r was taken at 10% as the company's balance sheet showed that production money was yielding that amount.

Adjustments were made in investment requirements for installed boiler capacity to reflect the better efficiency of the larger turbogenerator units. Operating labor costs were found to have an insignificant effect on the comparison and were omitted. An odd situation developed with respect to fuel. Being sludge and a waste product which required disposal if not burned, it had zero value.

The outcome of these manipulations showed conclusively that the unit size most appropriate to today's condition was a 4,000-kw unit. Both the next size smaller and the next size larger unit were found to be more costly in the long run. After two or three units of this size have been installed it would again be worth while to repeat the analysis under the then current conditions to see whether a different size should be adopted for succeeding units.

Summary

The range of uncertainties in planning an industrial power system can be narrowed and the planning placed on a more definite basis by evaluating future costs as well as immediate costs. The present-worth approach is a satisfactory way of doing this and of relating future costs to present costs in their proper order of magnitude. The ration I_p/I_d, the prebuilding cost to the cost of deferred construction, together with a knowledge of the cost of money and the expected life of the facility will guide planners in knowing when prebuilt facilities are economical. In complex problems a year-by-year analysis is needed.

The results of such an analysis will not be a final criterion for design. It is difficult to take into account all the future factors which may arise. In the first place, many contingencies cannot be foreseen. And even those which are foreseen may be difficult to express in dollar values. However, an analysis of this kind will improve the chances that an electrical system will attain minimum long term cost. It will place design on a definite and understandable basis and be a valuable guide to management. It is thus believed that management can be more readily induced to include the electrical design as one of the factors in overall production plant planning.

Discussion

Paul H. Jeynes (Public Service Electric and Gas Company, Newark, N. J.): This paper is of much more general application than its title suggests. Actually, it is an authoritative essay on some important aspects of the fundamental arithmetic of all economic analyses.

The following comments are intended to clarify a few minor details. They are of the nature of refinements, and not meant as substantial changes in the author's first-rate presentation.

Retirement Dispersion

The author assumes that the investment has a normal life of k years, which presumably can be forecast. In such case, his depreciation charge d is the sinking-fund annuity that will recover the investment (less net salvage) in k years.

Of course, it is impossible to forecast the actual life of a specific investment or unit of plant with accuracy, just as it is impossible to predict the life of any individual person. On the other hand, probable average life can be forecast with reasonable assurance, based on experience with groups of similar individuals or similar plant. It is characteristic of a group that some units serve for longer and some for shorter periods than the average life.

The resultant depreciation annuity is designated by the standard symbol A_d, where the beta superscript signifies that it is the statistically defensible group-basis figure. A_d is always greater than d. In other words, d without the beta is virtually always an underestimate of depreciation costs, because it is rarely true that all units of a group are retired simultaneously exactly at the predicted average life.

Tables of A_d for various types of dispersion and average lives have been calculated. A

rather comprehensive table for presentation as an AIEE technical paper is in preparation.

To estimate δd, it is necessary to estimate the type of dispersion. This is normally done as part of the life estimate. If no past experience is available on which to base the estimate of dispersion, it is much better to assume some reasonable type such as S_2 rather than to ignore this factor. Ignoring it amounts to firmly predicting no dispersion which is the least likely of all possible eventualities.

Replacing d in all equations by δd selected from a table, refines the estimate without introducing appreciable complication.

Present-Worth Factors

The author observes correctly that the present worth of a perpetual annuity is obtainable simply by dividing the annual charge by rate of return r. He proceeds to say that uniform annual operating differentials may be converted into "equivalent investments" by dividing by the fixed charge rate ρ or ρd. This seems to call for a little clarification.

The present worth of the annuity representing operating differentials is not found by dividing by ρ or ρd. Their present worth for one lifetime, k years, is found by multiplying by factor $A(=A_k, A_h, \text{or } A_L)$. $A=1/(r+d)$ i.e., the reciprocal of return plus depreciation. It may be obtained from standard tables of present worth of an annuity of $1.00, as the author indicates. For plant having dispersed retirements, as is always the actual case, $A=1/(r+\delta d)$.

The present worth of future operating differentials is not an investment, but a commitment; and it is by no means fictitious.

That is not what the author means by his term "equivalent investment," obtained by dividing $(e+i)$ I by ρ or ρd. He means specifically that the investment I' which, at carrying-charge rate ρ or ρd, would have total annual costs $(\rho I')$ equal to the operating differential $(e+i)$ I. I' is necessarily less than $Ax(e+i)I$. This equivalent investment is very commonly confused with the present worth of operating costs by those who attempt to "capitalize" annual expense.

Income Taxes

The author, in referring briefly to the effect of rapid depreciation on income taxes, says, "Its effect therefore is to make the true rate of interest (or dividends) on borrowed funds less than the apparent rate."

The view is technically defensible, in that, if some part of total capital is borrowed at zero interest without change in the rate of return on the other components, percentage return on the aggregate capital amount necessarily is reduced. But, practically, such a statement is misleading and invites trouble.

The rate of return r, as designated by the author, is the percentage that investors in the enterprise realize on their capital investment. The fact that some capital not contributed by investors commands zero return, does not directly effect the rate r paid to investors on the funds they do contribute.

The major effect of rapid depreciation is to reduce income taxes in the early years of an investment, at the expense of greater taxes later on. The equivalent level tax annuity is therefore a smaller percentage of first cost I. There is no direct effect on rate of return to investors r, any more than there would be by reason of larger "straight-line" reserves reinvested in the enterprise at zero return to the investor.

In fact, the larger capital ventured (equal to the investor's funds and reserves plus reinvested tax deferrals) to produce a given dollar of return to the owners, increases their risk to a degree, and (all other things being the same) might conceivably be expected to cause a small increase in the return demanded r.

To restate the situation in accounting terms, the effect of rapid depreciation for tax purposes is to decrease the rate of return on gross plant. But r is by definition the rate of return on net plant, and r is not directly affected.

The frequent need for year-by-year analyses, and the great advantage of answers in terms of present worth rather than break-even solutions, both of which the author points out, will be heartily agreed to by all having practical experience in this field.

W. H. Dickinson (Esso Research and Engineering Company, Linden, N. J.): The author is to be commended for his useful graphic methods of presenting his results in this thought-provoking paper. Most of the following comments are intended to indicate some points which may be helpful in applying the paper to individual problems.

The author indicates how to handle problems in which the differential operating costs are important. However, in many cases, the differential operating costs are negligible; for equations 20 and 21 make it comparatively easy to decide whether to spend money now for capacity to be required in the future.

Equation 20, $I_p/I_d=P_h$, covers the case when the facilities are not renewed until the end of their own economic life. This, I believe, applies generally to most generating distribution, and substation facilities. In effect, the equation merely states that the present worth of the deferred investment I_d at the required rate of return r is equal to the present investment I_p. Equation 25 shows how to modify equation 20 to take into account anticipated future price rises.

When the prebuilt and deferred facilities come to the end of their economic life at the same time, equation 21 $I_p/I_d=(P_h A_k \rho d)/A_k \rho$ applies. These equations require nothing more difficult than a calculation of

$S_n=(1+r)^n$, the compound interest factor (28)

From this

$d_n=r/(S_n-1)$, sinking fund depreciation (29)

$\frac{1}{A_n}=r+d_n$, the capital recovery factor (30)

$P_n=\frac{1}{S_n}$, the present worth factor (31)

In making calculations involving r, one must be careful to use the desired rate of return on investment after federal income taxes. The value of r to be used varies according to company policies but it usually depends on the risk being taken. The return rate is generally equal to that which can be obtained on similar investments of equivalent risk. A rate of return in the range of 10% to 20% would be common for industrial plant generation, distribution, and substation facilities. The lower percentages would be applicable when predictions about the future are expected to be reasonably accurate, and the higher when the predictions are not very certain.

The remaining comments have to do with the fixed charge factor.

$\rho=e+i+r+d$ (32)

It is suggested that calculations would be simpler if $r/(1-t)$, where t is the federal income tax rate, were substituted for the second and third terms i and r.

Regarding the last term d, the sinking-fund depreciation, the following is offered as a refinement, which, while more complicated, is believed to give more accurate results.

The method of treating depreciation in economic analyses has long been controversial. The facts are, however, that the Internal Revenue Service permits for income tax purposes, the use of several different types of depreciation, such as straight-line, sum-of-digits, or declining-balance, all of which have a higher incidence in the earlier years of an investment than does sinking-fund depreciation. This is advantageous as it permits earlier recovery of capital investment, and in effect increases the return on investment.

On the other hand, sinking-fund depreciation correctly recovers the investment from the standpoint of an economic analysis. Thus, it would appear that both income tax and sinking-fund depreciations should be used in making an economic analysis. In calculating the net income after taxes, one of the permitted income tax depreciations is used. This income tax depreciation is added to the net income to get what is called "cash-flow." Cash-flow is equal to the investment multiplied by the capital recovery factor $(r+d)$, d being the sinking-fund depreciation.

If we let d_t designate the income tax depreciation, then the foregoing considerations would require that $(d-td_t)/1-t)$ be substituted for d, the last term in equation 5. The derivation of this follows. To refer to the illustrative application in the paper for a 10% return, the value of ρ calculated by this method is about 10% lower.

Derivation of ρ

Let

- N = net income after taxes
- R = gross revenues
- O = operating costs and expenses not proportional to investment
- e = expenses proportional to investment, such as maintenance and property taxes
- I = total investment
- t = federal income tax rate
- d_t = equivalent uniform annual depreciation rate for income tax purposes
- d = sinking-fund depreciation rate
- r = required rate of return on investment after taxes
- c = cash-flow $= N+d_tI$

Then

$$N=(R-O-eI-d_tI)(1-t) \quad (33)$$

But

$$N+d_tI=(r+d)I \quad (34)$$

or substituting and collecting terms

$$I\left(e+\frac{r}{(1-t)}+\frac{(d-td_t)}{(1-t)}\right)=R-O \quad (35)$$

this gives,

$$I_p=R-O \quad (36)$$

where the fixed charge factor

$$p=e+\frac{r}{(1-t)}+\frac{(d-td_t)}{(1-t)} \quad (37)$$

replaces that used in the paper and given in equation 32.

D. Dalasta (Allis-Chalmers Manufacturing Company, Milwaukee, Wis.): While the author stresses that his application of long-range economic studies to industrial power systems must be supplemented by both engineering and management judgment, little is said of the many other factors which tend to reduce the value of the strict economic analysis. It is the intent of this discussion to cover some of these points. The author's own examples are used for the purpose.

To consider the example illustrated in Fig. 4, the economic comparison is based on the installation of a single large transformer at the present time or a smaller transformer now and an additional small transformer later. The additional investment for prebuilding is stated to be $17,800 while deferring construction would cost about $25,000. This gives a difference in transformer cost of about $7,500 and an Ip/Id ratio of about 0.7. However, some of the attendant factors, if ignored, would make the foregoing analysis quite misleading. The author mentions operating differences due to electrical losses in the larger transformer during the waiting period. Other items appear to be at least as important, among which are space requirements and installation cost of two transformers as compared to one. The increase in installation cost can readily be estimated but the space consideration varies widely with different applications. Also to be considered is that improved continuity of service at reduced capacity can be obtained with the two transformer installations. This brings into the picture the cost of primary and secondary switching facilities for the second transformer. For example, a single 1,200-ampere 13,800-volt circuit breaker would accommodate the secondary of the larger transformer while it would require two such breakers (at a cost of $7,000 to $10,000 each) to accommodate both transformers. Differential relays plus associated current transformers, if installed, would add another $2,000. Similarly, there would probably be primary switching facilities. It may be seen that service requirements will very often dictate the purchase of associated equipment minimizing greatly the consideration of transformer costs alone.

Much the same could be said concerning the example illustrating the selection of interrupting ratings for switchgear. Current-limiting reactors take valuable space, and because of their size they cannot be conveniently located in a metal-clad switchgear line-up. Thus, installation expense becomes important because of the metal-clad bus tie or cable connections required. Also, disconnecting devices might be required for isolation purposes. Generally, for the same voltage class, metal-clad switchgear units have essentially the same dimensions for all interrupting ratings. The actual current-limiting reactor and housing costs may sometimes be doubled if all allied factors are considered. The author's statement concerning the advantages of excess interrupting capacity in allowing the planning engineer greater freedom of design in the unforeseen future is certainly substantiated by the facts.

It is the intention of this discussion to supplement the author's excellent treatment of the application of economic studies to industrial power systems with the thought that the planner should search well to uncover all elements affecting the choice among alternative plans.

H. B. Backenstoss: By stating that this paper is of more general application than its title suggests, Mr. Jeynes has pointed to its most important aspect. It is indeed a development of relations which apply to economic analyses in whatever field they may be encountered. It was only by circumstance that the vehicle for their exposition was this particular subject of industrial power system planning.

Mr. Jeynes, a close student of depreciation, has done important work on retirement dispersion as it affects the amount of annual accrual necessary to recover original investment. It is, of course, essential to use the most accurate information available in the equations developed in this paper, which naturally includes a value for depreciation reflecting suitable dispersion whenever a group of items rather than a single unit is involved. It is to be noted that Mr. Jeynes, by proposing this refinement, supports the sinking fund method of determining actual depreciation cost for comparing alternative plans. This is significant, because a difficult fundamental problem which must often be met before any refinements whatever can be considered is to persuade users of straight-line depreciation of the basic error of that method, unless investment return is determined year by year on a declining investment base.

Mr. Jeynes's clarifying remarks on the subject of present-worth factors are welcome. Certainly this is a confused matter in economic studies. The confusion does not exist in reconciling the seemingly conflicting procedures mentioned in the paper for determining the present worth of annual charges, i.e., dividing by r in the case of a perpetual annuity and multiplying by $A_{b,n,r}$ in the case of a limited annuity. The annuity factor A is the reciprocal of $(r+d)$ as Mr. Jeynes notes, and d is zero in the case of a perpetual annuity, so these operations are equivalent. This procedure, and here is where the confusion arises, is distinctly one which does not "capitalize" annual expense. The investment equivalent of an annual expense annutiy is found by dividing the expense by the total fixed charge rate ρ or p_d, as the paper points out, and as Mr. Jeynes confirms. An annual charge extending for, say, b years must first be converted to an equivalent, and necessarily lower, annual value extending over the full equipment life of h years. It is done by multiplying by the factor A_b/A_h. The value of equivalent investment I' so obtained is thus related to annual operating expense by the same factor ρ as the actual capital investment is related to the annual fixed charge expense thereon.

Operating differentials which the author has in mind are typified by copper loss, core loss, labor differences, etc., rather than the term $(e+i)I$ which Mr. Jeynes mentions. His comments are equally valid regardless of the nature of the annual expense. It simply is not correct to capitalize by multiplying by the factor A.

In a communication to the author, Mr. Jeynes has suggested that the effect of rapid depreciation in reducing federal income taxes in the early years of an investment may conveniently be expressed as a reduction in tax rate in per cent of first cost. The author will readily concur in this and express regret if his statement describing its effect as a lowering of the true rate of return below the apparent rate left the impression that the value of r should be modified to suit. This would not be so, and r, the rate of return to investors, must also be the value of interest return for selecting present-worth and annuity factors. It should be mentioned in passing that the illustrative typical fixed charge rates presented in Table IV do not allow for federal tax savings, due to using a depreciation rate for tax purposes which exceeds sinking fund depreciation. Such was not their purpose.

Accordingly, Mr. Dickinson's comments on values to be used for depreciation are helpful in redeeming the author from his restricted coverage of this important subject. His equations eliminate federal tax as a separate item and incorporate it with those other items of total fixed charges which determine the tax. Inasmuch as the amount of depreciation allowed for tax purposes affects the tax actually paid, the factor t appears in the expression he develops for the depreciation component. This is a matter of personal preference but the author believes it is simple and direct to handle income tax separately from return and depreciation. When this is done, depreciation tax savings may readily be incorporated in the tax item by the use of the expression $i=t/1-t(r-BI-D)$, where D is the difference between d_t, the depreciation rate allowable for tax purposes, and d, the sinking-fund rate, i.e., (d_t-d). The term BI represents bond interest payment. In case part of capital funds are raised through bonded indebtedness, interest payments are, like depreciation, a tax-deductible expense. State income taxes, if applicable, can be added by a suitable, similar equation.

Although Mr. Dalasta refers to the many other factors which tend to reduce the value of the strict economic analysis, his examples actually constitute an effective buttress to the author's argument. No conclusion can be any stronger than its premises; it is consequently important to use the soundest and most complete factual data possible in solving any problem in planning. This was the author's purpose in presenting a method by which future costs might accurately be compared with present costs. By thus remov-

ing them from the area where they are considered solely by intuitive judgment, if at all, that area becomes less broad. Hence judgment becomes more effective. Mr. Dalasta presents an impressive list of side factors which bear on the economic choice of transformer size and which can and ought to be included in planning.

The author had hoped that there would be some discussion on the subject of price rise, which is merely touched upon in the paper. It is no doubt an exceedingly complicated subject on which there are many conflicting opinions. Much could be done in research to determine whether construction plans undertaken in the past would have been beneficially different had the planners foreseen the price situation which has subsequently developed. This does not mean research on the advantages of hindsight, which would be valueless, but rather on the question of whether it would have been wise to design facilities on the basis of an upwardly trending price level. The results would have a bearing on whether plans made today should anticipate further price rises. Should an affirmative answer be supported by the findings of such research, the result might well be a material impact on the pattern of capital expansion throughout American industry. If the author has stimulated thinking in this direction, this phase of the paper has served its purpose.

Rectifier Power Supply for a Modern Rod Mill

R. A. BUCHANAN
ASSOCIATE MEMBER AIEE

W. R. HODGSON
ASSOCIATE MEMBER AIEE

THE steel industry in the United States is a large user of d-c electric power. Since the use of d-c power for rolling mill drives is so widespread, it is natural to consider the application of ignitron rectifier equipment as the power supply for main mill drive service. Until approximately 20 years ago, the motor-generator set was used exclusively to supply the large amounts of d-c power for rolling mill service. Recently, however, ignitron rectifiers have found increasingly frequent use in the making of steel.

The ignitron rectifier provides an excellent means of conversion of alternating current to direct current because of its many ideal features, such as low maintenance, high efficiency, and low installation cost. It gives the steel industry a tool, the many inherent desirable features of which will add momentum to the already increasing number of applications for main mill drive service. The ignitron rectifier is being purchased almost exclusively for 250-volt supply for auxiliary motors and is increasing for 600- or 700-volt main drive supplies.

The purpose of this paper is to present to the industry the ignitron rectifier of the continuously evacuated metal-enclosed type as applied in a modern rod mill and the salient features of this equipment as applicable to the main mill drives of our modern up-to-date steel mills. The growing trend toward rectification is evidenced by installation at a large midwestern combination rod mill.

Paper 56-44, recommended by the AIEE Industrial Power Rectifiers Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Winter General Meeting, New York, N. Y., January 30-February 2, 1956. Manuscript submitted October 21, 1955; made available for printing November 17, 1955.

R. A. BUCHANAN and W. R. HODGSON are with the Westinghouse Electric Corporation, East Pittsburgh, Pa.

Over-All Station Equipment

This is the first application of ignitron rectifiers in a rod mill as main drive power supply. The mill is a 25-stand combination rod mill. A total of three ignitron rectifiers is utilized to furnish power to the 10 mill motors. The no. 1 rectifier feeds a total of 3,600 horsepower (hp) of d-c motors and is rated 2,920 kw. The no. 2 rectifier feeds a total of 4,000 hp of d-c motors and is also rated 2,920 kw. The no. 1 and no. 2 rectifiers are exact duplicates. The no. 3 rectifier has a total of 5,900 hp connected to its bus, all of which are not energized at the same time. The type of operation and the mill schedule govern the number of hp actually fed by the no. 3 rectifier, which is rated 3,500 kw.

The mill motors are 700 volts direct current with adjustable speed by shunt field control. The mill equipment is designed to enable successful operation of the high-speed mill with four strands of bar.

The control and switchgear equipment does not include any bus tie switches or circuit breakers, and the three bus sections and their respective supply rectifiers will operate independently at all times. The direct voltage on any of the three bus sections is independently variable and not dependent on the other bus sections. The entire rectifier substation with the exception of the transformer is built in metal-enclosed construction and located in the motor room.

It should be noted from the simplified schematic diagram of Fig. 1 that the three d-c busses are separate and that the power supply on each bus is identical except for the increased kilowatt capacity required by one of the units.

Each of the rectifier units consists of an outdoor rectifier transformer, a 6-pole metal-enclosed drawout anode circuit breaker, a metal-enclosed 12-tank pumped ignitron rectifier assembly, a d-c control cubicle, and a drawout single-pole cathode breaker. In addition, a master control cubicle, an a-c breaker and pulpit control are common to all of the rectifiers.

In Fig. 2 the compact arrangement of the metal-enclosed cubicles in the motor room and the outdoor rectifier transformers may be seen. The metal-enclosed construction of the rectifiers affords ease of maintenance, simple installation, and neater station layout. In most cases this affords easy planning of the installation in advance for existing structures.

D-C Conversion Equipment

The rectifier power transformers are the oil-immersed self-cooled outdoor type. Two of the rectifier transformers are connected delta on the primary and one is connected wye on the primary. This affords the benefits derived from unbalanced 12-phase operation as far as a reduction in harmonics is concerned in the a-c system. Since split busses are used on the d-c side, the units will operate as 6-phase units as far as d-c loads are concerned.

The transformer secondaries will be of the conventional 6-phase double-wye connection. Incorporated in the same tank with each of the three rectifier power transformers is a regulating autotransformer with a tap-changing underload mechanism. This mechanism is utilized to vary the voltage under load from 700 volts d-c to 410 volts.

The primaries of the 3-rectifier power transformers will be controlled together from one 11,500-volt circuit breaker, as shown in the simplified schematic

Fig. 1. Power supply diagram for modern rod mill

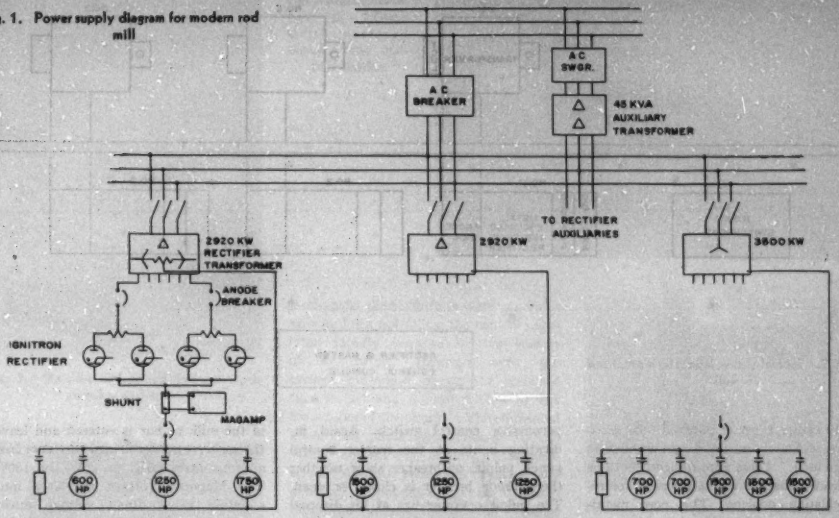


diagram of Fig. 1, and are designed for a 3-phase 60-cycle system. Each rectifier power transformer will have a 3-pole no-load primary disconnect switch in the transformer high-voltage terminal chamber, so that any transformer may be disconnected during checking of individual rectifier units or other emergency operation. Bushing current transformers are provided to afford individual a-c overload protection of each of the rectifier power transformers.

The anode breaker is a metal-enclosed 6-pole high-speed electrically operated drawout air circuit breaker, designed with individual tripping and closing of each pole. There are no mechanical or electrical connections between the six poles as they are entirely isolated from each other. The poles can be individually adjusted.

Each pole unit is equipped with a high-speed tripping mechanism. All pole units are independently trip-free in any part of the closing operation, thus allowing the separate clearing of arc-back faults. The use of 6-pole automatic-reclosing high-speed anode switchgear maintains the continuity of service. The control circuits are arranged to permit the closing of any combination of open poles without disturbing those which are closed.

The drawout breaker cell incorporates a counterbalanced vertical-lift sliding door which allows a reduction in aisle space over the conventional type of door construction. The circuit breaker is conveniently withdrawn from the cubicle and, with the wheels which can be changed for the direction of motion, can easily be handled by one person. Fig. 3 shows the breaker being racked out of the cubicle, and the convenient overhead door structure. Fig. 4 shows the breaker racked out of the cell with one boot barrier removed to show the construction of the breaker.

The ignitron rectifier is metal-enclosed and close-coupled to the anode circuit breaker on one side and the d-c control on the other. This facilitates interconnections. Fig. 5 shows a front view of the rectifier with all doors closed, while Fig. 6 shows a front view of the rectifier with all doors open showing the interior equipment. Fig. 7 is a rear view of the cubicle with doors open, showing interior equipment. The cubicle contains 12 pumped ignitron rectifier tanks, anode balance reactors, phase-shifting circuits, and a water-to-water heat exchanger for rectifier cooling purposes. Each of the main power supply metal-enclosed ignitron rectifiers is in a factory-assembled dead-front metal-enclosed cubicle and of matching dimensions and appearance with the switchgear breaker and control cubicles, which, when installed together, form a co-ordinated rectifier unit substation.

With the advent of the metal-enclosed rectifier, it is possible to have completely metal-enclosed unit construction in which drawout anode and cathode switchgear can be close-coupled to the metal-enclosed rectifier unit. Since all internal power control circuits and connections are built into the complete unit, the metal-enclosed type of construction will result in a very substantial reduction in the installation time and expense, amounting to approximately 25%, as compared to installation of other types of conversion equipment.

With this construction, it is necessary only to install the anode and negative connections from the outdoor transformer to the anode circuit breaker, the raw water-pipe connections to the rectifier unit, the positive and negative connections to the d-c bus. This arrangement eliminates the necessity of the purchaser's providing anode connections from the anode circuit breaker to the rectifier, control connections between the rectifier and firing cubicle, cathode connections from the rectifier to the cathode circuit breaker, and negative connections to and mounting of negative disconnecting switch. It also eliminates the external control connections between the anode circuit breaker, the rectifier, and the cathode switchgear.

In addition to the reduction in installation cost and time, the metal-enclosed unit will require less building volume because of its compactness. The elimination of necessary protective screens with the metal-enclosed units means smaller substation buildings at considerable cost saving to the ultimate customer.

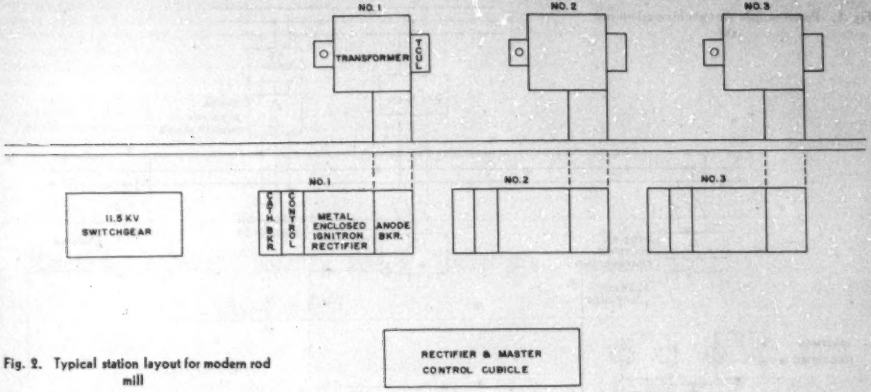


Fig. 2. Typical station layout for modern rod mill

In earlier types of pumped tube rectifiers, separately mounted heat exchangers were used. These were designed both as a heat exchanger and a reservoir for recirculating coolant. The new metal-enclosed pumped rectifier has a compartment for the heat exchanger and water reservoir, as shown in Fig. 7. The heat exchanger can be easily cleaned by removing top and bottom header plates. The water will drain off into the convenient drain pan. The shells can then be cleaned by running cleaning rods down from above. The vertical position facilitates flushing the tubes after rodding. The recirculating pump is submerged, driven from a vertically mounted motor on top. Leakage around the shaft of such a pump simply returns water to the low-pressure reservoir. The metal-enclosed rectifier and associated metal-clad equipment provide a pleasing appearance in a substation or motor room.

Mill Operation

The first step in normal operating procedure is to close the anode circuit breakers. On the anode breaker cubicle, individual indicating lights indicate when each pole is closed; while on the master control cubicle a red light indicates only when all six poles are closed. When the lockout relay is reset and all components are working satisfactorily, the master control switch can be operated and the a-c breaker closed. Satisfactory closing of the a-c breaker will be indicated on the master control unit by a red light. Having completed the necessary sequence checks, the main control panel operator can then turn the cathode breaker control over to the pulpit operator through a permissive control switch. Again indicating lights at the master control panel, pulpit, or breaker show whether the cathode breaker is closed or open. The pulpit operator has at his disposal complete control of the motors he is to operate whether it be operating at nominal d-c voltages or jogging at low d-c voltages. However, in each case he must properly select the motors and reference voltage (other than jogging) before starting, or the control on the excitation circuits of the ignitron rectifier will not release. It is possible for the operator to select "jog" or "hold" operating voltage. This is accomplished simply by opening a contact, and short-circuiting the resistance in the field circuit of the exciter, for jogging and stopping the motor-operated reference voltage rheostat in mid-travel for hold.

Special Features

Incorporated in the over-all installation are a number of special features, one of which is operation to the maximum possible power factor regardless of operating voltage, except during jogging and initial start-up. This feature is accomplished by incorporating in each case of the rectifier power transformers a regulating autotransformer with taps connected to a load tap changer. A total of 16 steps on the tap changer provides equal percentage steps between the nominal d-c voltages of 410 to 700 volts.

Once the operating voltage has been determined, the tap changer selecter switch is turned to the automatic position. In this position the tap changer mechanism is controlled by an automatic control relay. During normal operation of the mill, as bar is entered and leaves the various stands the load may vary from approximately 25% to 125% or 150%. The Magamp voltage regulator must maintain rated direct voltage during these load swings. If the load swings to high peaks, the phase shift circuit may be unable to advance far enough to maintain rated voltage and a tap change will be necessary. If the load drops excessively the rectifier will be phased back by the regulator to maintain the operating direct voltage. This of course means a reduction in power factor (pf) of approximately 1 per cent for each per cent of voltage reduction by phase control.

The automatic control relay controlling the tap changer mechanism assures operation at the highest possible pf by utilizing a minimum amount of voltage control by phase retarding of the rectifier.

The control relay receives its intelligence from the bias current in the phase shift circuit of the excitation circuits. If the load swing is such that the direct voltage attempts either to fall below or rise above operating voltage, the regulator will cause corrective action by phase-controlling the rectifier. However, if the phase control current shifts above or below a preselected value, the control relay will operate, causing a tap change to provide the proper operating voltage with minimum phase retard.

The lower and upper limits of the relay can be so adjusted to operate as near as possible to maximum pf and at the same time to insure a minimum number of tap changer operations. Typical operating calculations indicate that with the control relay operating the tap-changing-under-load mechanism, a possible pf saving may be obtained of 4% to 8%

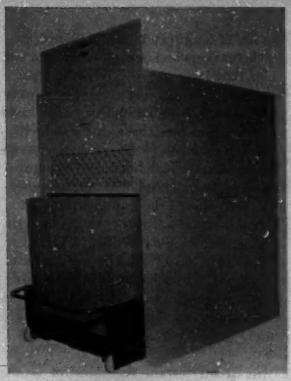


Fig. 3. Six-pole high-speed drawout anode breaker and cell

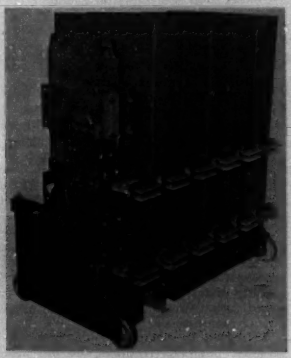
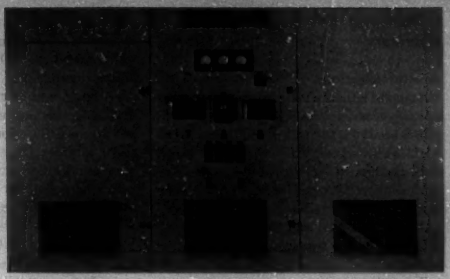


Fig. 4. Six-pole high-speed anode breaker assembly with one boot barrier removed

Fig. 5 (right). Front view of 12-tank metal-enclosed pumped tube ignitron rectifier



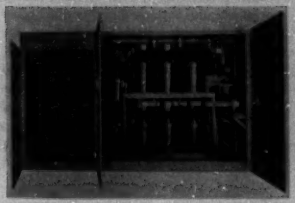


Fig. 6. Interior front view of 12-tank metal-enclosed pumped tube ignitron rectifier

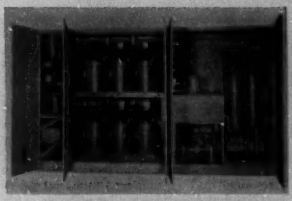


Fig. 7. Interior rear view of 12-tank metal-enclosed pumped tube ignitron rectifier

at 125% load and 7% to 8% at 25% load, depending upon upper and lower control relay settings. To prevent hunting and excessive tap changer operations due to momentary changes in supply voltages, a delay circuit has been incorporated in the control circuit. This time-delay circuit is adjusted for the best operation of the mill and the least amount of maintenance of the tap changer.

Another special feature of this rectifier power supply is the complete control of the rectifier direct voltage through phase control with the use of the saturable reactor-type excitation circuit. In the earlier installations, in order to obtain full phase control it was standard practice to use mechanical phase shifters. These mechanical phase shifters were generally slow and did not follow the control regulator rapidly because of the inertia of the rotor. However, with the present excitation circuit the response time is fast and is limited only by the inductance of the circuit. The additional amount of phase shift required for full voltage control was obtained by the addition of another phase shift reactor and auxiliary components to the standard excitation circuit. With full voltage control, the operator has complete control for friction and jogging loads and for accelerating the mill to normal direct voltage.

It is well known that rectifiers cannot absorb power during the time the motors are regenerating. Normally this power is absorbed by other loads connected to the d-c bus. However, if other loads will not absorb this regenerated power and the bus is exclusively supplied by rectifiers, means of absorbing the power must be provided. This is done through the addition of a regenerative resistor connected across the bus. The resistor is automatically controlled so that it does not dissipate energy under normal load conditions.

For a main drive supply it is important that the direct voltage be closely regulated. Each of the main power supply ignitron rectifiers will be controlled by a Magamp type of voltage regulator, which will maintain essentially constant voltage on each bus throughout the load range. Since the inherent direct voltage from a rectifier is directly proportional to the alternating voltage, provisions are made in the transformer design to prevent any temporary dips in the a-c system from being reflected on the d-c bus. Consequently, the rectifier power transformer must be designed for above-rated direct voltage in order that voltage will be available for subnormal alternating voltage conditions. The capacity and range of the ignitron rectifier transformers and their corresponding Magamp regulator are sufficient for the rated direct voltage to be obtained from friction to 125% loads, with only 95% alternating voltage applied to the transformer primary windings. The three voltage regulators as well as the individual motor speed regulators receive their intelligence from a common reference so that the whole system will respond equally to any reference voltage change. Also, each rectifier Magamp regulator will have individual adjusting rheostats to enable operation with the same or different voltages on the three bus sections. The individual voltage-adjusting rheostats will vary the intelligence to the speed regulators of the motors from the respective bus sections, so that speed regulator settings will vary proportionally to the Magamp regulator settings.

Summary

The selection of the continuously evacuated ignitron tube for d-c conversion is evidenced by over 5,000,000 kw in conversion service today. With the inherent high overload capacity of the pumped tube rectifier, it is well suited for the duty cycle imposed by steel-mill service. The low long-time maintenance, the availability for service, and the high operating efficiency are factors leading to the application of the pumped tube rectifiers for mill service.

The present-day trend with rectifier power supplies is toward higher and higher d-c bus voltages. Although voltages for mill auxiliaries are fixed at 250 volts d-c, main-drive voltages of 600 are usual and 700 volts are becoming common. The economic factors favoring the use of rectifiers at 700 volts are compelling. In addition, the efficiency advantage of a rectifier over other types of conversion equipment for main mill drives is more pronounced at 700 volts, amounting to approximately 8% at full load. With the large number of kilowatts being utilized, this advantage means a real saving in dollars for the mill operator.

In addition to these advantages, the user usually saves on the initial cost when purchasing rectifier equipment. The installation is approximately 75% of that for rotating conversion equipment. The rectifier being a static piece of equipment does not require the heavy foundation required for rotating machinery. Likewise, the need for ventilating equipment is practically eliminated as the rectifier equipment is water-cooled. In addition to these advantages, the metal-enclosed unit offers the user all the best features of the pumped ignitron in a package unit.

The reliability of the rectifier has been proved during the past 13 years for steel-mill operation. During the past few years the reliability has improved with the use of high-speed automatic-reclosing anode breakers and control circuits to the point that it no longer affects the choice between rectifiers and motor-generator sets. Along with reliability, the high overloads which rectifiers can stand without any damage to its components as compared to the high brush and commutator wear experienced with motor-generator sets should be appreciated. Another important advantage of the rectifier is the lower maintenance cost. A number of auxiliaries must be maintained, such as the cooling and vacuum systems; however, it is generally established that the cost of maintaining a rectifier is approximately 60% that of a motor-generator set.

The trend toward rectifiers for steel-mill drives results in part to their many advantages, and it is reasonable to expect a gradual spread in the application of rectifier equipment to main mill drive service.

Discussion

W. E. Miller (General Electric Company, Schenectady, N. Y.): The advantages of rectifiers for continuously running mills, infrequently started and stopped, have been recognized for some years. The introduction of the compact, metal-enclosed, pumpless rectifier unit substation further enchances the future of rectifiers as a power supply for main roll drives because they provide additional reduction in installation cost, cut installation time, increase safety to personnel and equipment, and make low rectifier maintenance even lower.

Actually, the first installation of rectifier main-mill drive systems of the type described by the authors (utilizing regulating autotransformers, packaged metal-enclosed rectifier substations with magnetic amplifier control systems) was placed in operation on McLouth Steel's 6-stand hot-strip finishing mill in 1954. This installation, utilizing six such individual ignitron rectifier power supply systems, has been completely described in papers presented before the Association of Iron and Steel Engineers during early 1954 and early 1955.

Since that time several such equipments have been sold for use as power supply systems for rod and skelp mill drives. The first such ignitron rectifier equipments actually to be placed in service as main drive power supply for a rod mill are located in the Pittsburgh area and have been in operation producing rod since December 1955. Rolling schedules are such that this mill will not be required to operate at reduced direct voltage so that the load ratio-regulating autotransformer feature was not required on the three rectifier equipments supplied. However, this feature is incorporated in a mill in a south eastern state.

Both the hot-strip finishing mill and the rod mill installations have proved beyond doubt the reliability and the desirable performance features of this type of ignitron rectifier drive system. It is noted, however, that the installations in operation differ to some extent from the installation described by the authors. The authors point out that the installation described utilizes autotransformers over a range of 700 volts to 410 volts with only 16 steps. The authors also point out that the direct voltage changes instantaneously with the alternating voltage. Thus, with only 16 steps on the regulating autotransformer, the transient effect of one bus section changing one step with reference to another will be greater than the transient impact speed drop of a motor for either 3- or 4-strand operation. It has been our practice to supply 32-step tap changers to ensure that the effect of a relative change in one tap position between bus sections would have less stretching effect on a rod than on (or would) a rod strand leaving and entering a stand.

All four of the main-drive installations referred to in this discussion have not only been of the packaged metal-enclosed type but also were supplied with pumpless or sealed tanks. One of the major considerations leading to the selection of this type of rectifier equipment on recent main roll drives has been that these equipments eliminate another rectifier maintenance problem, the vacuum-pumping system. No longer is it necessary for the user to worry even about the vacuum, nor about the quality of the water supplied to the pumps.

R. A. Buchanan and **W. R. Hodgson**: Mr. Miller's discussion is of considerable interest since it reports of further acceptance of ignitron rectifiers in the steel industry.

There appears to be little basis for concern about the performance of the 16-step regulator. Under normal operation with an armature voltage regulator and a speed regulator on the motor, any disturbance in speed is almost negligible. A 1-step tap change is not as severe as the impact speed drop of a motor when a strand enters the rolls.

We were well aware of the 32-position tap changer mechanism but, in discussion with the user, the 16-position tap changer seemed more attractive. With this mechanism the number of tap changes is considerably reduced with a subsequent decrease in transformer maintenance.

In regard to the comments on metal-enclosed pumpless rectifiers, we are also aware of the fact that elimination of vacuum pumping system simplifies, to a slight degree, the over-all routine maintenance of a rectifier installation. However, the maintenance of the excitation circuits, the cooling system, and the switchgear equipment requires in each case more attention than the vacuum pumping system. The choice of the continuously evacuated ignitron rectifiers was that of the user. He has had in service over 12 years the continuously pumped type of tubes and has found that the vacuum system requires only very minor maintenance attention. He is well satisfied with the ruggedness of the continuously pumped type of tubes, their proved ability to withstand heavy overloads, and their greater over-all long-time economy, since there is no periodic tube replacement or major reconditioning required. These advantages are incorporated with those of package-type construction in the metal-enclosed pumped ignitron as described in the paper. Should the user desire a sealed tube design, these are available in high-capacity ratings for steel mill service.

Analysis and Performance of a Valve-Controlled Hydraulic Servomechanism

DAVID V. STALLARD
ASSOCIATE MEMBER AIEE

VALVE-CONTROLLED hydraulic servomechanisms are an important class of control systems because of their widespread use in fields requiring high performance and light weight, particularly for military aircraft. Unfortunately, hydraulic servos are not widely understood, as manifested by the scarcity of good technical papers and textbook material on the subject.

Hall[1] and Shearer[2] have contributed noteworthy papers on valve-controlled hydraulic servos. Hall developed an equivalent electric network for the valve and motor; Shearer presented a rigorous, thorough derivation of the dynamic operation of a valve-controlled ram with a combined load of mass, viscous damping, and external force. It would be helpful to extend Shearer's analysis to include the effect of mechanical compliance in a rotary motor and to derive equations for the velocity constant and torque constant of the complete control system. Data from an experimental system to illustrate the theory would also be helpful.

The tendency of the electrical engineer to analyze in terms of impedance parameters and networks can be useful in understanding hydraulic servos. This paper, which is essentially a condensation of reference 3, analyzes the dynamic behavior of a single-stage valve and single-vane gearless rotary motor, Fig. 1, with inertia load. An equivalent mechanical network is presented with parameters to account for oil compressibility, line elastance, mechanical compliance, motor leakage, etc. In addition, the velocity constant and torque constant of the complete positional control system, Fig. 2, are derived. The analysis could be extended readily to servos with 2-stage valves and other types of hydraulic motors.

In the experimental system presented as an example, the calculated values of motor resonant frequency, motor damping ratio, gain-crossover frequency, and torque constant agreed well with the measured values; the gain-crossover frequency ω_{pc} is defined as the lowest frequency at which th loop ratio B/E is unity and is decreasing with frequency. The large mechanical compliance of the single-vane gearless servomotor lowered its resonant frequency considerably. It is likely that unsuspected compliances seriously affect the performance of many hydraulic control systems.

Analysis of Valve, Servomotor, and System

In this analysis, the positional control system comprises an amplifier, a single-stage control valve (operating from constant pressure), a single-vane rotary motor, Fig. 1, connected directly to an inertia load, and position feedback from the load to the amplifier, as shown in Fig. 2.

Significance of Motor Radian Displacement

An important parameter of a positive-displacement motor is its radian displacement D_r. If negligible leakage is assumed, the radian displacement is the volume of oil that enters (and leaves) the machine during one radian of rotation. Hence the motor flow rate Q_m is

$$Q_m = D_r\Omega \tag{1}$$

where Ω is the motor speed. If the friction is negligible, then the hydraulic power input equals the mechanical power output

$$P_mQ_m = T_m\Omega \tag{2}$$

Equation 1 may be substituted into equation 2 to give an expression for motor torque

$$T_m = D_rP_m \tag{3}$$

From the foregoing equations, it is apparent that the size, required flow rate, and output torque of a machine are proportional to its radian displacement. Furthermore, it is shown later that a machine with a large displacement has a low oil compliance and a high resonant frequency in the ideal case. Thus, the choice of a suitable radian displacement tends to be a compromise between various conflicting requirements.

Valve Performance

In Appendix I, an analysis is given for a 4-way control valve with moderate underlap (and therefore moderate null leakage flow rate Q_L) operating from a constant system pressure P. Because the differential pressure P_m across the motor is usually small, it is convenient to write the expression for motor flow rate Q_m as

$$Q_m = k_xX - \left|\frac{\partial Q_m}{\partial P_m}\right| P_m \tag{4}$$

For small values of stroke X and P_m, the partial derivative may be related to the leakage flow rate

$$Q_m = k_xX - \left(\frac{Q_L}{2P}\right)P_m \tag{5}$$

To apply equation 5 to the motor, it is divided by D_r, and T_m/D_r is substituted for P_m

$$\frac{Q_m}{D_r} = \frac{k_x}{D_r}X - \frac{Q_L}{2P}\left(\frac{1}{D_r^2}\right)T_m \tag{6}$$

$$= \Omega_f - F_rT_m \tag{7}$$

The term Ω_f is equal to the steady-state velocity of an unloaded motor for a constant valve stroke X. The parameter F_r accounts for the slight decrease of motor flow rate Q_m with motor pressure P_m.

In a single-stage valve, a spring-restrained, electromagnetic torque motor develops a magnetic force and a steady-state stroke that are both proportional to the input differential current. As shown in Appendix I and reference 4, the effect of the steady-state flow forces on the valve piston is equivalent to an extra centering spring with a spring constant of $0.43\ wP_s$, where P_s is the total pressure drop across the valve orifices. For small system errors, P_m is small and P_s is almost equal to P. This spring constant may be added to that of the valve torque motor if the valve pistons do not have force compensation. Hence, it appears justifiable to use a transfer function X/I from valve input current to valve stroke, and a separate transfer function from valve stroke to motor velocity.

Paper 56-197, recommended by the AIEE Feedback-Control Systems Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Winter General Meeting, New York, N. Y., January 30-February 3, 1956. Manuscript submitted October 5, 1955; made available for printing December 12, 1955.

DAVID V. STALLARD is with the Massachusetts Institute of Technology, Cambridge, Mass.

The research reported in this paper was made possible through the support extended the Massachusetts Institute of Technology, Servomechanisms Laboratory, by the U. S. Air Force (Weapons Guidance Laboratory, Wright Air Development Center), under *Contract No. AF33(616)-2039, M.I.T. Project No. D.I.C. 7138.*

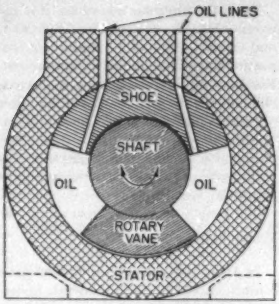


Fig. 1. Rotary hydraulic motor

Dynamic Performance of Motor

In the steady state, a constant valve stroke X produces a constant motor velocity that is equal to $\Omega_f \equiv k_x X/D_r$. The dynamic operation is much more complicated because of the effects of load inertia, leakage, and the compliances arising from oil compressibility, line elastance, and torsional deflections of the motor parts. A very helpful tool for understanding these effects is the equivalent mechanical network.

Equation 65 of Appendix II is the Laplace transform of the dynamic equation relating the load angle θ to valve stroke X

$$\Omega_f \equiv \frac{k_x X}{D_r} = s\theta + (F_v + F_m)Js^2\theta + (c_o + c_m + c_L)Js^3\theta \quad (8)$$

Because the load is an inertia J, the transformed torque T_m acting on it is equal to $Js^2\theta$, which may be factored out of equation 8 to give

$$\Omega_f \equiv \frac{k_x}{D_r}X = T_m\left[\frac{1}{Js} + F_v + F_m + (c_o + c_m + c_L)s\right] \quad (9)$$

The bracketed terms in equation 9 may be regarded as mechanical impedances through which the motor torque flows. They are connected in series in the equivalent mechanical network shown in Fig. 3.

The motor output torque T_m is equal to the reaction torque exerted by the motor against its mounting. To account for torsional deflections of the mounting, the equivalent mechanical network includes a mounting compliance c_b in series with the load inertia, although it was not included in the analysis.

The electrical analogue of the mechanical network is shown in Fig. 4, with voltage analogous to velocity and current analogous to torque. This circuit is equivalent to the familiar resistance-inductance-capacitance low-pass filter.

The transfer function relating load velocity $s\theta$ to valve strokes X is readily found from equation 8

$$\frac{s\theta}{X} = \frac{k_x/D_r}{1 + FJs + cJs^2} \quad (10)$$

$$\frac{s\theta}{X} = \frac{k_x/D_r}{1 + 2\zeta\frac{s}{\omega_n} + \left(\frac{s}{\omega_n}\right)^2} \quad (11)$$

where F represents $F_v + F_m$, c represents $c_l + c_o + c_m$. The resonant frequency ω_n is given by

$$\omega_n = \frac{1}{\sqrt{Jc}} \quad (12)$$

and the damping ratio is

$$\zeta = \frac{\omega_n}{2}FJ = \frac{F}{2}\sqrt{\frac{J}{c}} \quad (13)$$

From the preceding derivation, it is apparent that the frequency response of

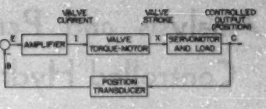


Fig. 2. Elementary valve-controlled hydraulic servomechanism

load velocity for a sinusoidal valve stroke of constant amplitude is constant for frequencies well below ω_n, peaks near ω_n, and falls off at 40 decibels per decade (20 decilogs per decade) at higher frequencies. At the resonant frequency ω_n, the phase angle of equation 11 is -90 degrees.

The mechanical resonance at ω_n is an interchange of energy between the load inertia and the total compliance, and is generally the most troublesome obstacle in obtaining a high gain-crossover frequency in the system. To maximize the resonant frequency, it is well to evaluate and minimize the various compliances.

If the rotor of the machine in Fig. 1 is centered, then the oil compliance is given by

Nomenclature

The basic units employed are: pound (lb), force, inch (in), second (sec), and radian (rad). From $F = ma$, the unit of mass is taken as lb-sec^2 per in.

B = primary feedback
C = controlled variable
c = total compliance of motor (equation 66), rad per lb-in. (Compliance is the ratio of deformation angle to motor torque.)
c_L = line compliance (equation 57), rad per lb-in
c_m = mechanical compliance (equation 62), rad per lb-in
c_o = oil compliance (equation 58), rad per lb-in
D_r = radian displacement of motor, in^3
E = actuating signal (error)
F = total reciprocal-damping factor (equation 66), rad per lb-in-sec
F_m = motor reciprocal-damping factor (equation 59), rad per lb-in-sec
F_v = valve reciprocal-damping factor (equation 44), rad per lb-in-sec
I = signal current to control valve, milliamperes
J = moment of inertia of load, lb-in-sec^2
K = gain constant of loop ratio (equation 17), rad per sec
k_a = line elastance, i.e., ratio of volumetric increase of line to pressure increase, in^3 per lb
K_T = torque constant of servo, rad per lb-in
K_v = velocity constant of servo, sec^{-1}
k_x = flow sensitivity of valve (equation 35), in^2 per sec
L_m = motor leakage coefficient (equation 54), in^5 per lb-sec
P = system pressure, lb per in^2
P_1 = pressure drop across valve orifice no. 1, lb per in^2
P_2 = pressure drop across valve orifice no. 2, lb per in^2
P_m = pressure drop across motor, lb per in^2
Q_1 = volumetric flow rate through valve orifice no. 1, in^3 per sec
Q_2 = volumetric flow rate through valve orifice no. 2, in^3 per sec
Q_L = total leakage flow rate through valve at zero stroke, in^3 per sec
Q_m = volumetric flow rate through motor, in^3 per sec
Q_{mo} = flow rate through motor at zero motor pressure, (equation 35), in^3 per sec
R = reference input
s = Laplace transform variable
T_L = torque disturbance applied to servo, lb-in
T_m = torque developed by motor, lb-in
U = underlap of valve port, in
V_m = total static volume of oil between one port and motor vane with vane at center position, in^3
w = width of valve port, in
X = stroke of valve, in
β = bulk modulus of oil, (equation 51), lb per in^2
ζ = damping ratio of motor (equation 69)
θ = angle of load inertia, measured clockwise, rad
ρ = mass density of oil, lb-in^{-4}-sec^2
ϕ = angle of motor vane, measured clockwise, rad
Ω_f = equivalent input velocity of valve, (equation 8), rad per sec
ω_c = asymptote gain-crossover frequency of loop ratio, rad per sec
ω_{gc} = gain-crossover frequency of loop ratio, rad per sec
ω_n = resonant frequency of motor (equation 68), rad per sec

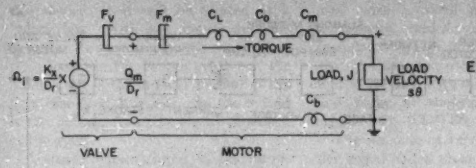


Fig. 3. Equivalent mechanical network from valve stroke to load velocity

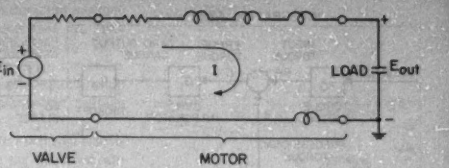


Fig. 4. Electrical analogue of the mechanical network

$$c_t = \frac{V_m}{2\beta D_r^2} \tag{14}$$

V_m is the oil volume on one side of the motor and is slightly more than D_r times the maximum angular swing from center. The bulk modulus β of the oil may be assumed to be 250,000 lb per in^2, unless the oil is excessively gassy and the system pressure is low. Reference 5 gives measured values of 4.0×10^{-6} in^3 per lb for oil compressibility, which is the reciprocal of bulk modulus.

The line compliance c_L is caused by the elasticity of the lines connecting the valve and motor, and is given by

$$c_L = \frac{k_s}{2D_r^2} \tag{15}$$

where k_s is the ratio of the volume change of one line to the pressure change in that line. The line compliance probably will be large if a gauge or transducer of the Bourdon type is connected to each motor line, particularly if the motor displacement is small.

For a machine of the single-vane type, as in Fig. 1, c_m is caused by deformation of the vane fasteners and by twisting of the shaft between the vane and load. The compliance of a simple cylindrical shaft is inversely proportional to the fourth power of its diameter, and tends to be appreciable. In an optimum servomotor, the mechanical compliance would be very much smaller than the oil compliance.

It is apparent that the significance of each unwanted compliance may be found by comparing it with the oil compliance, which is usually irreducible. In the ideal situation, the only compliance would be the oil compliance. From equation 12 the resonant frequency would be

$$\omega_n = \frac{1}{\sqrt{Jc_t}} \tag{16}$$

which is equivalent to the equation given by Hall[1] and others.

Closed-Loop Response

In the positional control system of Fig. 2, a position signal B from a transducer is fed back from the load to a summing point in the valve amplifier. Because the valve-driven motor is essentially a velocity source at low frequencies, the loop ratio B/E has one integration. If the gain is adjusted for a gain-crossover frequency that is very much lower than the mechanical resonant frequency, then the mechanical resonance of the motor will not be important. If it is necessary to achieve a high gain-crossover frequency in the position loop, then the mechanical resonance poses a difficult stability problem. When the loop gain is increased, the system tends to oscillate at a frequency near the mechanical resonant frequency, even though the gain-crossover frequency may be less than half the mechanical resonant frequency. In this case, the system must be compensated so that the resonance is damped and/or raised to a higher frequency. There are a number of ways[2,11] of compensating a hydraulic servo and, unfortunately, a thorough treatment of this topic would require a separate paper. Reference 11 describes feedback compensation from an accelerometer mounted on the load. In some cases, it is possible to insert into the valve amplifier a tuned electric network which cancels the motor resonance with an antiresonance.

It is of interest to consider the gain-crossover frequency, velocity constant, and torque constant of a simple hydraulic servo, and to show how these three parameters are interrelated. The simplest case to consider is that of a servo with a single positional loop and no network compensation, as shown in Fig. 2. From equation 11 and equation 48 the loop ratio is seen to be

$$\frac{B}{E} = \frac{KG_v}{s\left[1+2\zeta\frac{s}{\omega_n}+\left(\frac{s}{\omega_n}\right)^2\right]} \tag{17}$$

where G_v is the frequency-dependent part of the transfer function X/I for the valve. For a single-stage valve, G_v would be the reciprocal of a quadratic and would have a value of unity at zero frequency (cf. equation 48). The parameter K is a gain constant with the dimensions of rad per sec.

The gain-crossover frequency ω_{gc} has been defined as the lowest frequency at which the loop ratio is unity and is decreasing with frequency. The actual gain-crossover ω_{gc} is usually nearly equal to the asymptote gain-crossover frequency ω_c at which the asymptote of the loop-ratio decibel magnitude, plotted versus log frequency, is unity.

It is apparent from equation 17 that the asymptote gain-crossover frequency ω_c is the gain parameter K. Furthermore, the velocity constant K_v of the system is also equal to K. To a good approximation, the actual gain-crossover frequency is equal to ω_c and, therefore, equal to K and K_v. It is convenient to relate these parameters to ω_{gc} for two reasons: 1. ω_{gc} is readily determined by plotting frequency-response measurements of the closed loop on the log modulus-angle plane; and 2. ω_{gc} is a rather good measure of the rise time of the system transient response to a step input.

If an integral network with a ratio α of break frequencies were added to the amplifier in the system and ω_{gc} were not changed, then K_v would be approximately $\alpha\omega_{gc}$. Finally, it is worth while to relate the velocity constant to two velocity parameters in the mechanical network in Fig. 3

$$K_v = \frac{\Omega_o}{E} = \frac{\Omega_i}{E} \tag{18}$$

For systems in which there may be an appreciable static load torque (e.g., from friction or an unbalanced rotary load), it is well to determine K_T, which is the ratio T_L/E of load torque to system error. The torque constant is readily related to the velocity constant by a simple analysis. However, for a system encountering rapidly changing load torques (e.g., gun-recoil loads), a more refined analysis would be needed to determine the transient system error.

A steady load torque T_L is applied to the directly driven load, and is opposed by an equal motor torque T_m developed from the system error E. This load may be illustrated as in Fig. 3 by adding a torque source T_L to the mechanical network

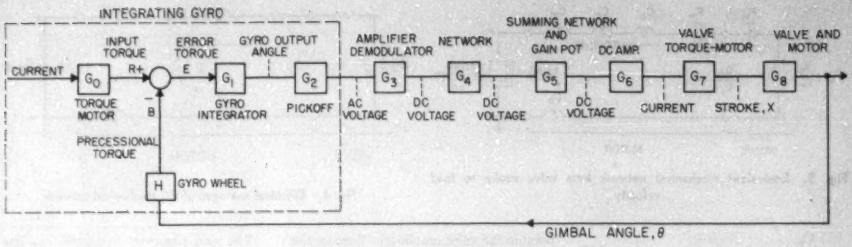


Fig. 5. Block diagram of stabilization loop

diagram in parallel with the load inertia. (In the electrical analogue in Fig. 4, a current source would be connected in parallel with the capacitor.) From equation 18 it is apparent that the valve which has a constant stroke develops an Ω_t

$$\Omega_t = K_v E \tag{19}$$

Because of the feedback loop, T_m equals the load torque T_L and there is no output velocity Ω_0. An application of the final value theorem (of Laplace Transform theory) shows that, in the steady state, equation 63 becomes

$$\Omega_t = T_m(F_v + F_m) \tag{20}$$

Equations 19 and 20 may be combined to eliminate Ω_t and form K_T

$$K_T = \frac{T_L}{E} = \frac{T_m}{E} = \frac{K_v}{F_v + F_m} \tag{21}$$

For a given valve and supply pressure, the parameter F_v is proportional to the leakage flow rate through the valve orifices. For many systems, F_v is likely to be negligibly small compared to F_m, which is proportional to the leakage coefficient L_m of the motor. Since the leakage through the small motor clearances is laminar with a small Reynold's number, it depends on the oil viscosity which varies appreciably with temperature. With increasing temperature, L_m and F_m increase and K_T decreases. Hence, the torque constant K_T is actually not constant for a hydraulic system, but this term is retained from the classical servo literature for want of a better one with equal acceptance.

It is apparent from equation 21 that increasing F_v or F_m to increase the damping ratio ζ of the motor resonance will lower the torque constant. The use of an integral network to increase K_T and K_v will add to the system response a relatively slow transient that may be objectionable.

Example of a Valve-Controlled Servo

The system chosen to illustrate the preceding theory is the gyro-stabilization loop, of an experimental, hydraulic antenna drive[3] for an aircraft application. The gyro-stabilization loop is actually a rate loop (with an output velocity proportional to reference input) rather than a positional control system. However, the experimental system has a loop ratio B/E with one integration, as does a positional control system, and is therefore suitable as an example.

Description of System

The gyro-stabilization loop consists mainly of an amplifier, a compensating network, a fast single-stage control valve, a single-vane hydraulic motor (Fig. 1) without gearing, a flywheel to simulate the antenna inertia, and an integrating gyroscope (gyro) connected as shown in the block diagram of Fig. 5.

The control valve is a single-stage *DACL SS* valve[6] developed by the Dynamic Analysis and Control Laboratory of the Massachusetts Institute of Technology. It is a 4-way valve with very fast response and force compensation for stroking forces. Its permanent-magnet torque motor directly strokes two identical, adjustable spool pistons of 0.250-in diameter through a maximum stroke of ±0.007 in. Experimentally it was found necessary to adjust the valve pistons for a leakage flow rate of about 0.96 in³ per sec to deliver enough flow to the motor at full stroke. Maximum flow rate was 8.1 in³ per sec for counterclockwise motor rotation, and 9.8 in³ per sec for clockwise rotation. Both the linearity and force compensation were rather poor, probably because of wear in this particular valve. To minimize line compliance and oil volume, the valve was connected directly to the motor by a special steel manifold. The valve was supplied with *MIL-0-5606* oil at 1,000 pounds per in² from a pump system that included a 1-gallon accumulator to reduce dynamic pressure fluctuations.

The hydraulic motor is a specially rebuilt Hydromotor model *3-1*, originally manufactured by the Hydromotor Division of the Bonnot Company, Canton, Ohio. As shown in Fig. 1, this machine has a single rotor vane attached to the shaft, and a single shoe fixed to the stator. Two flat end plates (not shown) bolted to the stator carry the ball bearings and oil seals. Oil entering the right side of the motor, Fig. 1, causes it to turn clockwise and expel oil from the left side, and vice versa. Static pressure in the motor is about one half of system pressure. The maximum angular travel from center is ±65 degrees.

From the standpoint of oil consumption and weight, this motor is rather large, for it has a radian displacement of 1.89 in³ and a weight of about 18.8 lbs. For the valve flow rates given in the foregoing, the maximum counterclockwise velocity of the motor is 8.1/1.89 or 4.3 rad per sec, and the maximum clockwise velocity of the motor is 9.8/1.89 or 5.18 rad per sec.

The load driven directly by the motor is a steel flywheel to simulate the load inertia. It has a diameter of 11.75 in, a weight of about 22 lbs, a weight moment of inertia of 400 lb-in², and a mass moment of inertia of 1.04 lb-in-sec². From equation 3, the maximum motor torque at stall is 1,000 (1.89) or 1,890 lb-in, which gives the load an acceleration of 1,890/1.04 or about 1,820 rad per sec².

Mounted on the flywheel is a *HIG-5* integrating gyro,[7] model *GG1C-1*, manumanufactured by Minneapolis-Honeywell Regulator Company. It is a floated, single-degree-of-freedom unit with a torque generator for forcing precession and a signal generator for measuring the output angle. Space does not permit a complete description of the operation of

this device, so it will be treated as a "black box," with two inputs and one output. As shown in Fig. 5, one input is the current in the torque generator, the other input is the angular motion of the gyro case (and flywheel), and the output is the voltage from the signal generator of the gyro.

A steady current input to the torque generator produces a steady torque and a steady rate of increase of the voltage from the signal generator. An incremental change in the angle of the case (about the flywheel axis) produces a proportional change in the voltage from the signal generator; hence, at low frequencies, the gyro acts like a position transducer. Similary, a steady velocity of the gyro case produces a constant rate of increase of the output voltage. The transfer functions of the gyro are given in equations 22 through 25.

In the experimental system shown in Fig. 5, a constant input signal to the torque generator produces a constant torque R, which is opposed by a precessional torque B proportional to the speed of the flywheel. In the steady state, the output signal of the gyro, the valve stroke, and the servomotor speed are constant. The precessional torque of the gyro equals and opposes the torque of the torque generator. Hence, the gyro acts as the summing point of the feedback loop.

The electronic portion of the system includes: an amplifier demodulator for the output signal of the gyro; a compensating network; a direct-coupled amplifier to drive the valve torque-motor; a modulator and a-c amplifier to drive the gyro torque generator; and a temperature regulator to control the gyro temperature. The output stage of the torque motor amplifier has two *6AK6* pentodes with very high plate resistance; hence, the inductive phase lag of output current behind the grid signal is negligible.

Frequency Response of Valve and Servomotor

Because the servomotor is the most limiting element in the system, the open-loop frequency response of the valve and motor for sinusoidal signals was measured rather carefully. An electromechanical sine drive furnished a sinusoidal voltage of constant amplitude and variable frequency to the amplifier driving the valve; hence, at low frequencies, the valve stroke and motor velocity had a constant rms value. A very small, linear-transformer transducer was used to measure the valve stroke dynamically. The sinusoidal velocity of the flywheel was measured with a small permanent-magnet tachometer, which generated an unmodulated voltage proportional to velocity. A voltmeter, oscilloscope, and phase unit of the sine drive measured amplitude and phase of the tachometer voltage signal.

The response of valve stroke to amplifier signal had a phase lag that was approximately proportional to frequency and was about 18 degrees at 56 cycles per second (cps). The resonant frequency of the torque motor is about 250 cps.

The frequency response $s\theta/I$ of load velocity for sinusoidal amplifier signals is shown in Fig. 6 and includes the phase lag of the valve torque motor. The velocity amplitude at low frequencies was 0.78 rad per sec rms.

It is worth while to correlate the observed resonant frequency with the calculated value. In Appendix II it is shown that at the resonant frequency the sinusoidal motor velocity lags behind the valve stroke by 90 degrees. This frequency was experimentally determined as 56 cps, by observing on an oscilloscope a circular Lissajous pattern between the valve-stroke signal and the motor-velocity signal.

The resonant frequency may be calculated from the theory in Appendix II. From equation 58, the oil compliance c_0 is calculated as 1.252×10^{-4} rad per lb-in, using a value of 250,000 lb. per in^2 for the bulk modulus β. This value of β may be slightly high. To measure the mechanical compliance, the motor was rotated until the vane contacted the shoe, the flywheel was loaded by a spring scale, and the static deflections of the flywheel under various loads were measured with a dial indicator. In this way, the mechanical compliance c_m was experimentally determined as 5.0×10^{-4} rad per lb.-in. It is surprising that this compliance is about four times the oil compliance. Forty per cent of the mechanical compliance definitely was caused by torsional deflection of the motor shaft, and the remainder was attributed to rocking of the vane on the shaft.

With these values of c_0 and c_m, and a load inertia J of 1.04 lb.-in-sec^2, the resonant frequency is calculated from equation 12 as 392 rad per sec, corresponding to 62.4 cps. The agreement between this calculated resonant frequency and the measured value of 56 cps is unusually good for such a calculation. The 11-per-cent error probably is not all attributable to the inevitable experimental errors. It may be due partly to some subtle effects, such as compliance in the motor bearings or moderate gassiness of the oil.

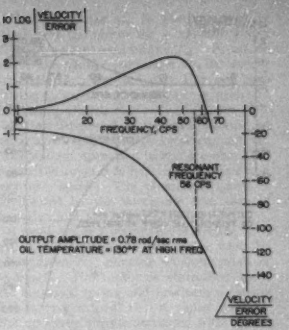


Fig. 6. Frequency response of valve and motor

From the frequency response data in Fig. 6, the damping ratio ζ at the oil temperature of 130 degrees Fahrenheit is inferred as approximately 0.3. The theoretical damping ratio may be calculated from Appendix II. From static leakage tests of the motor, its leakage coefficient L_m at 130 degrees Fahrenheit was found to be 4.1×10^{-9} in^5 per lb-sec. From equation 59, the corresponding motor reciprocal-damping factor F_m is determined as 1.15×10^{-4} rad per lb-in-sec. The valve reciprocal-damping factor F_v is calculated from equation 43 as 0.134×10^{-4} rad per lb-in-sec using a value of 0.96 in^5 per sec for valve leakage. With these values of F_m and F_v, and the previously computed compliances, ζ was calculated from equation 13 as about 0.26. The 13-per-cent error between 0.26 and the experimental value of 0.3 is partly caused by friction of the O-rings in the motor. Because the motor leakage increases with temperature, ζ also will increase.

Closed-Loop Response of System

In preliminary tests without a compensating network, it was found that the closed-loop response tended to peak sharply or exhibit instability near the mechanical resonant frequency. A resonant lead network, therefore, was inserted into the block marked G_4, as shown in Fig. 5. Theoretically, the resonant lead network cancels the mechanical resonance with an antiresonance at 56 cps and substitutes a moderate resonance at 125 cps. This technique requires matching and is not good at all temperatures because of the variation of motor damping with temperature, but it is useful for research purposes.

The transfer functions in Fig. 4 are:

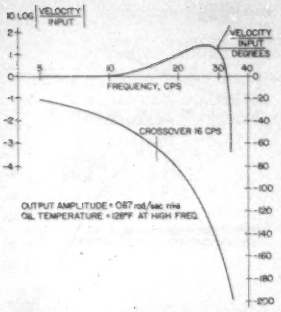

10 LOG VELOCITY INPUT
VELOCITY INPUT DEGREES
FREQUENCY, CPS
CROSSOVER 16 CPS
OUTPUT AMPLITUDE = 0.67 rad/sec rms
OIL TEMPERATURE = 128°F AT HIGH FREQ.

Fig. 7. Closed-loop frequency response of system

gyro torque motor

$$G_9 = 500 \frac{\text{dyne-centimeters}}{\text{milliamperes}} \tag{22}$$

gyro wheel

$$H = 10^3 s \left(\frac{\text{dyne-centimeter-sec}}{\text{rad}} \right) \tag{23}$$

gyro integrator

$$G_1 = \frac{10^{-3}}{s\left(1+\frac{s}{314}\right)} \frac{\text{rad}}{\text{dyne-centimeter-sec}} \tag{24}$$

gyro pickoff

$$G_2 = 32 \frac{\text{volt a-c}}{\text{rad}} \tag{25}$$

amplifier demodulator

$$G_3 \cong 29.7 \frac{\text{volt d-c}}{\text{volt a-c}} \tag{26}$$

network (resonant lead controller)

$$G_4 = 0.186 \frac{1+2(0.3)\frac{s}{352}+\left(\frac{s}{352}\right)^2}{1+2(0.25)\frac{s}{785}+\left(\frac{s}{785}\right)^2} \frac{\text{volts}}{\text{volts}} \tag{27}$$

summing network and gain potentiometer

$$G_5 = 0.070 \frac{\text{volts}}{\text{volts}} \tag{28}$$

d-c amplifier

$$G_6 = 34.3 \frac{\text{milliamperes}}{\text{volts d-c}} \tag{29}$$

valve torque motor

$$G_7 \cong \frac{0.3}{1+2(0.2)\frac{s}{1,570}+\left(\frac{s}{1,570}\right)^2} \frac{\text{mils}}{\text{milliamperes}} \tag{30}$$

valve and motor

$$G_8 = \frac{0.793}{s\left(1+2(0.3)\frac{s}{352}+\left(\frac{s}{352}\right)^2\right)} \frac{\text{rad}}{\text{mil}} \tag{31}$$

(One mil is 0.001 in.)

As in measuring the open-loop response, an electromechanical sine drive was used to furnish a sinusoidal voltage with constant amplitude and variable frequency. This signal drove the amplifier connected to the gyro torque motor, so as to call for a constant-amplitude sinusoidal velocity of the system. The same tachometer generated a signal proportional to output velocity, the amplitude and phase of which were measured as before. Fig. 6 is a plot of the frequency response of flywheel velocity relative to the amplifier input signal. At low frequencies the velocity amplitude was 0.67 rad per sec rms. The oil temperature at high frequencies was about 128 degrees Fahrenheit. Fig. 7 shows the closed-loop frequency response of the system.

From a plot of the data on the log-magnitude angle plane (Nichols chart), it was seen that the frequency of unity loop gain was about 16 cps at which the apparent open-loop phase lag was 126 degrees. The most convenient way to check these data against open-loop data is to add all the phase shifts at 16 cps around the loop:

Valve current to motor velocity (Fig. 6)	− 20°
Integration for velocity to position	− 90°
Total gyro phase shift = $-\tan^{-1}(16/50)$ = −17.8°	− 17.8°
Demodulator phase shift	− 7.5°
Calculated network phase shift	+ 7.1°
Total phase shift around loop	−128.2°

This calculated value of −128.2 degrees checks with the apparent value of −126 degrees quite well.

Friction of the motor is important, because it limits the angular accuracy of the stabilization loop. A measure of the stiffness of the stabilization loop to torque disturbances is the torque constant, for which an analytical expression was derived in equation 21. Torque constant is usually defined as the ratio of static load torque to static error. Actually, it varies with temperature.

Experimentally, the torque constant was determined by loading the flywheel with a spring scale and measuring the gyro output voltage. This voltage was proportional to the gyro gimbal angle, which was equal to the angular deflection of the flywheel. These data were reduced to a graph of load torque versus angular deflection. For torques below 220 lb-in the plot was close to a straight line with a slope of 71.5 lb-in per millirad, which is taken as the torque constant K_T at the mean test temperature of 99 degrees Fahrenheit. For larger torques, the data points were somewhat scattered. As expected, the torque constant was found to decrease with temperature because the motor leakage increased.

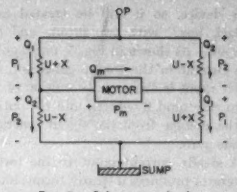


Fig. 8. Schematic of valve

It is well to check the measured torque constant against the calculated value with the use of test data and equation 21. Interpolation of motor leakage data at 99 degrees Fahrenheit gives a motor leakage coefficient L_m of 3.36×10^{-3} in^4 per lb-sec, which corresponds to a motor reciprocal-damping-factor F_m of 0.94×10^{-3} rad per lb-in-sec. As shown previously, the value of the valve reciprocal-damping-factor F_v is 0.134×10^{-3} rad per lb-in-sec. The radian crossover frequency corresponding to 16 cps is 100.5 rad per sec. Hence, by equation 21, the torque constant is $100.5/1.074\times10^{-3}$, which is 93.5×10^3 lb-in per rad or 93.5 lb-in per millirad. This value is 31 per cent higher than the experimental value of 71.5 lb-in per millirad and represents a fairly good check for this type of calculation.

This high value of torque constant resulted in a very low dead band due to friction and apparently contributed to the very wide range of smooth speeds. The ratio of maximum speed to minimum smooth speed was 3,700 or greater.

Conclusions

From the analysis and example, it is concluded that the dynamic behavior of a valve-controlled servomotor with inertia load may be analyzed accurately with linear transfer functions. Understanding is enhanced by the use of simple parameters and an equivalent mechanical network. The velocity constant and torque constant are readily calculated from the gain crossover frequency and other parameters.

Good agreement between measured and calculated values of motor resonant frequency and damping ratio in the example resulted from including mechanical com-

pliance and motor leakage at operating temperature in the calculations. In many hydraulic systems, unsuspected compliances probably cause the actual resonant frequency to be much lower than the frequency calculated from oil compressibility alone.

Appendix I. Analysis of a Single-Stage 4-Way Control Valve

The static behavior of the control valve may easily be examined first without reference to the dynamic behavior. A single-stage 4-way valve will be analyzed. It has four variable orifices connected to the pressure supply P, motor and sump, in bridge fashion, as shown in Fig. 8. All orifices have a port width w and are varied simultaneously in length by the valve torque motor. When a slight underlap U is assumed, two orifices have a length of $U+X$, while the other two have an opening of $U-X$. From Bernoulli's equation[9] and a discharge coefficient of about 0.6, the flow rates are:

$$Q_1 = 0.6(U+X)w\sqrt{\frac{2P_1}{\rho}}$$

$$= 0.6(U+X)w\sqrt{\frac{P-P_m}{\rho}} \quad (32)$$

$$Q_2 = 0.6(U-X)w\sqrt{\frac{2P_2}{\rho}}$$

$$= 0.6(U-X)w\sqrt{\frac{P+P_m}{\rho}} \quad (33)$$

$$Q_m = Q_1 - Q_2 = 0.6\frac{w}{\sqrt{\rho}}[(U+X)\sqrt{P-P_m} - (U-X)\sqrt{P+P_m}] \quad (34)$$

For small loop errors, the motor pressure drop P_m will be small because the required acceleration will be small. Therefore, the motor flow rate Q_m at zero motor pressure P_m is of interest. For strokes X that exceed the small underlap U, the flow Q_2 is zero and Q_m is simply

$$Q_{m0} = 0.6X\sqrt{\frac{P}{\rho}} \equiv k_x X \quad (35)$$

Hence it is apparent that for constant supply pressure and zero motor pressure, the motor flow rate is proportional to stroke. Experimental data on well-built valves confirm this linearity up to a large stroke at which the flow rate begins to approach a maximum because of saturation. Reference 9 shows complete performance curves of normalized motor flow versus normalized motor pressure for various valve strokes.

The motor flow rate at small values of motor pressure may be written as a linear approximation

$$Q_m \cong Q_{m0} - \left|\frac{\partial Q_m}{\partial P_m}\right| P_m \quad (36)$$

where both Q_{m0} and the derivative are evaluated at $P_m = 0$. From equation 34 the partial derivative in equation 36 is evaluated

$$\frac{\partial Q_m}{\partial P_m} = -\frac{0.6w}{2\sqrt{\rho}}\left(\frac{U+X}{\sqrt{P-P_m}} + \frac{U-X}{\sqrt{P+P_m}}\right) \quad (37)$$

This derivative is to be evaluated at zero motor pressure and a small value of stroke X. Operation generally occurs at small values of stroke except in slewing. It is apparent from equation 37 that the derivative is constant from a stroke of $-U$ to a stroke of $+U$. Beyond these values one or the other pair of orifices is closed completely. For this small-stroke condition

$$\frac{\partial Q_m}{\partial P_m} = -0.6wU\frac{1}{\sqrt{\rho}}\frac{1}{P}$$

$$= -\frac{1}{P}\left(0.6wU\sqrt{\frac{P}{\rho}}\right) \quad (38)$$

To calculate the derivative readily from experimental data, it is desirable to evaluate it in terms of the leakage flow. From Fig. 8, it is apparent that the leakage flow to sump at zero stroke is

$$Q_L = 2Q_1 = 2(0.6)wU\sqrt{\frac{2P_1}{\rho}}$$

$$= 2\left((0.6)wU\sqrt{\frac{P}{\rho}}\right) \quad (39)$$

Equation 39 may be substituted into equation 38 to give

$$\frac{\partial Q_m}{\partial P_m} = -\frac{Q_L}{2P} \quad (40)$$

For the normal case of a slightly underlapped valve, this value will be small. Now, equations 40 and 35 may be substituted into equation 36 to find the valve equation

$$Q_m = k_x X - \frac{Q_L}{2P}P_m \quad (41)$$

Two general conclusions can be drawn for the static behavior of a slightly underlapped 4-way valve.

1. At zero motor pressure, the motor flow rate is linear with stroke up to saturation.
2. At small values of stroke and motor pressure, the motor flow is that of a slightly imperfect flow source.

To apply equation 41 directly to the motor, it may be divided by D_r, and T_m/D_r may be substituted for P_m

$$\frac{Q_m}{D_r} = \frac{k_x}{D_r}X - \frac{Q_L}{2P}\frac{1}{D_r^2}T_m \quad (42)$$

It is convenient to define two quantities

$$\Omega_l \equiv \frac{k_x}{D_r}X \quad (43)$$

$$F_v \equiv \frac{Q_L}{2PD_r^2} \quad (44)$$

With these new parameters, equation 42 becomes

$$\frac{Q_m}{D_r} = \Omega_l - F_v T_m \quad (45)$$

The parameter Ω_l is the steady-state velocity of the motor for a constant valve stroke X and no load torque T_m. F_v will be added to a similar parameter F_m derived in Appendix II. The fact that F_v is not constant is unimportant, because it is generally small compared to F_m.

At low frequencies, the combination of motor and valve acts like a velocity source. Thus smooth operation in the presence of a reasonable amount of friction is possible at a very low speed. The maximum speed is determined by the maximum flow rate of the valve and the motor displacement.

A single-stage 4-way valve is stroked by a fast electromagnetic torque motor which is driven from two pentode vacuum tubes in push-pull. The motor is restrained by a stiff torsional spring and produces a torque that is proportional to the signal current. The magnetic torque increases with stroke because two of the four air gaps are being closed. The motor then has effectively a negative magnetic spring constant k_m, the magnitude of which must be subtracted from the mechanical spring constant k_t to find the net spring constant $k_t - k_m$.

If the valve has zero supply pressure and low sticking friction, it will have a stroke proportional to signal current. For a sinusoidally varying signal current, the total inertia and damping at the motor armature will be significant. There will be a resonance at a high frequency.

Now if the valve is operated at rated supply pressure with its two motor ports connected, oil will flow through the orifices and produce forces on the valve piston. Static flow forces on a valve spool inevitably result from the momentum of the oil jets at the orifices and always act to close the valve. For zero load pressure, the flow force is proportional to the orifice jet velocity and to the flow rate which, in turn, is proportional to stroke. Both the jet velocity and the flow rate are proportional to the square root of orifice pressure drop. Reference 4 gives this equation for the total static flow force

$$F_s = 0.43A_p P_v = 0.43wXP_v \quad (46)$$

where

A_p = the area of the port opening
P_v = the total pressure drop across the valve

Because this flow force is proportional to stroke and tends to center the spool, it is equivalent to the force of a spring with spring constant k_f

$$k_f = 0.43wP_v \quad (47)$$

Because P_m across a motor with inertia load is usually small, P_v is nearly equal to P. The net effective spring constant for the valve is $k_t - k_m + k_f$. Besides static flow forces, there are also dynamic flow forces, which tend to damp the valve motion.

Unless there are large eddy current or stiction forces acting on the valve armature, the transfer function from valve current I to valve stroke X is a quadratic

$$\frac{X}{I} = \frac{k_i}{1 + 2\zeta'\frac{s}{\omega_v} + \left(\frac{s}{\omega_v}\right)^2} \quad (48)$$

where

k_i = a static gain constant
ω_v = the armature resonant frequency

Appendix II. Valve-Motor Transfer Function

The transfer function $s\theta/X$ from the valve stroke X to the load velocity $s\theta$ will be derived.[1,2] The dynamic performance of the motor may be analyzed by the equation of continuity[10] in fluid mechanics, i.e., the principle of conservation of mass. In Fig. 2, the motor may be considered to be rotating clockwise at mid-position, with a vane angle ϕ measured clockwise from an arbitrary reference. Oil of mass density ρ flows from the valve into the right port with a volumetric flow rate Q_m at a pressure P_1. The pressure drop across the motor is $P_m = P_1 - P_2$. The line is assumed to have appreciable compliance (e.g., from a Bourdon-tube transducer or gauge) with a ratio k_s of volume increase to pressure increase. A control volume V_m is defined as the static volume of oil between the right valve port and the motor vane at mid-position. The equation of continuity may be stated as

$$\text{mass flow-rate into } V_m - \text{mass flow-rate out of } V_m = \text{rate of increase of mass in } V_m \quad (49)$$

Substituting into equation 49

$$\rho Q_m - \rho D_r\left(\frac{d\phi}{dt}\right) - \rho Q_L - \rho k_s\left(\frac{dP_1}{dt}\right) = \frac{d}{dt}(\rho V_m) = V_m\left(\frac{d\rho}{dt}\right) \quad (50)$$

In equation 50, the first term is the mass flow rate into V_m, the second term is the mass rate leaving V_m to push the vane clockwise, the third term is the leakage mass rate flowing from V_m past the vane, and the fourth term is the mass rate that swells the volume of the compliant line. The term on the right side is the rate of increase of mass in the fixed volume V_m, resulting from the compression of the oil, which has finite bulk modulus β.

$$\beta \equiv \frac{dP_1}{d\rho/\rho} \quad (51)$$

An equation similar to equation 50 could be written for the left side of the motor, which has a pressure P_2. It is assumed that pressure change dP_1 on the right is accompanied by a negative pressure change dP_2 on the left. Hence a change dP_m in the motor pressure drop can be written as

$$dP_m = dP_1 - dP_2 = 2dP_1 \quad (52)$$

From equations 51 and 52, the expression on the right side of equation 50 may be rewritten

$$V_m\left(\frac{d\rho}{dt}\right) \equiv \frac{\rho V_m}{\beta}\left(\frac{dP_1}{dt}\right) = \frac{\rho V_m}{2\beta}\left(\frac{dP_m}{dt}\right) \quad (53)$$

The third term in equation 50 is the laminar leakage rate past the vane, and may be rewritten as

$$Q_L = L_m P_m \quad (54)$$

Further, it is convenient to relate motor pressure to motor torque

$$dP_m = \frac{dT_m}{D_r} \quad (55)$$

Now, by using equations 52 through 55, equation 50 may be rewritten as

$$\frac{Q_m}{D_r} - \frac{d\phi}{dt} - \frac{L_m}{D_r^2}(T_m) - \frac{k_s}{2D_r^2}\left(\frac{dT_m}{dt}\right) = \frac{V_m}{2\beta D_r^2}\left(\frac{dT_m}{dt}\right) \quad (56)$$

It is now convenient to define certain expressions in equation 56 as parameters

$$c_L \equiv \frac{k_s}{2D_r^2} \quad (57)$$

$$c_0 \equiv \frac{V_m}{2\beta D_r^2} \quad (58)$$

$$F_m \equiv \frac{L_m}{D_r^2} \quad (59)$$

Compliance has the dimensions of angle/torque, and may be thought of as an inverse spring constant. It is convenient to use F_m, which is the ratio of a velocity loss to the motor torque, because F_m may be added later to a similar parameter F_v for the valve. Equations 45, 57, 58, and 59 may be substituted into equation 56 to give

$$\frac{k_s}{D_r}X - F_v T_m - \frac{d\phi}{dt} - F_m T_m - c_L\left(\frac{dT_m}{dt}\right) = c_0\left(\frac{dT_m}{dt}\right) \quad (60)$$

Since a transfer function is desired, the Laplace transform of equation 60 is formed and rearranged

$$\frac{k_s}{D_r}X = s\phi + (F_v + F_m)T_m + (c_0 + c_L)sT_m \quad (61)$$

Expression for ϕ and T_m now will be found. The angle of the motor vane is ϕ, and the angle of the inertia load is defined as θ, both angles being positive in the clockwise direction. It is assumed that an appreciable compliance exists between vane and shaft, and another compliance exists in the shaft from vane to load. These two mechanical compliances are lumped together as c_m. The load inertia J is so much greater than the rotor inertia that the latter may be neglected. Then the twist angle $\phi - \theta$ is

$$\phi - \theta = c_m T_m \quad (62)$$

Equations 43 and 62 may be substituted into equation 61 to give

$$\Omega_t \equiv \frac{k_s}{D_r}X = s\theta + (F_v + F_m)T_m + (c_0 + c_L + c_m)sT_m \quad (63)$$

For a load of inertia J

$$T_m = Js^2\theta \quad (64)$$

Equation 64 may be substituted into equation 63 to give

$$\Omega_t \equiv \frac{k_s}{D_r}X = s\theta + (F_v + F_m)Js^2\theta + (c_0 + c_L + c_m)Js^3\theta \quad (65)$$

Finally, a solution is found for the transfer function $s\theta/X$, which is the ratio of transformed load velocity to transformed valve stroke

$$\frac{s\theta}{X} = \frac{k_s/D_r}{1 + FJs + cJs^2} = \frac{k_s/D_r}{1 + 2\zeta\frac{s}{\omega_n} + \left(\frac{s}{\omega_n}\right)^2} \quad (66)$$

where F is the sum $F_v + F_m$, and c is the sum $c_m + c_0 + c_L$. The denominator of equation 66 is the familiar quadratic

$$1 + FJs + cJs^2 = 1 + 2\zeta\frac{s}{\omega_n} + \left(\frac{s}{\omega_n}\right)^2 \quad (67)$$

Hence

$$\omega_n = \frac{1}{\sqrt{Jc}} = \frac{1}{\sqrt{J(c_m + c_0 + c_L)}} \quad (68)$$

and

$$\zeta = \frac{\omega_n}{2}FJ = \frac{F}{2}\sqrt{\frac{J}{c}} \quad (69)$$

The comprehension of this analysis may be improved by reference to the equivalent mechanical network of the valve and motor, as shown in Fig. 3. It is derivable from equation 65 if the motor torque $Js^2\theta$ is factored out

$$\Omega_t \equiv \frac{k_s}{D_r}X = T_m\left[\frac{1}{Js} + F_v + F_m + (c_L + c_0 + c_m)s\right] \quad (70)$$

The bracketed terms in equation 70 may be viewed as mechanical impedances. The motor torque flows through each impedance. It may be shown that this torque is equal to the reaction torque exerted by the motor against its mount. Hence, in Fig. 3 a mounting compliance c_b is included in series with the motor inertia, although it was not included in the analysis.

For those who prefer electrical networks, the electrical analogue is shown in Fig. 4, with voltage analogous to velocity and current analogous to torque. This circuit will be recognized as the familiar resistance-inductance-capacitance low-pass filter.

From the preceding derivation it may be seen that the frequency response of load velocity for a sinusoidal valve stroke of constant amplitude is constant for frequencies well below ω_n, peaks near ω_n, and falls off at 40 decibels per decade at higher frequencies. At ω_n, the phase angle of equation 66 is -90 degrees. The resonance at ω_n is due to energy interchange between the load inertia and the system compliances, and is the most troublesome aspect of the stability problem.

References

1. Application of Frequency-Analysis Techniques to Hydraulic Control Systems, A. C. Hall. *Transactions*, American Society of Mechanical Engineers, New York, N. Y., vol. 76, no. 8, Nov. 1954, p. 1245.

2. Dynamic Characteristics of Valve-Controlled Hydraulic Servomotors, J. L. Shearer. *ASME Paper No. 53-A-147, Ibid.*, Dec. 1953.

3. Performance of an Experimental Hydraulic Antenna Drive, D. V. Stallard. *Report No. 7138-R-12*, Massachusetts Institute of Technology, Cambridge, Mass., Sept. 1955.

4. Contributions to Hydraulic Control—1. Steady-State Axial Forces on Control-Valve Pistons, S. Y. Lee, J. F. Blackburn. *Transactions*,

American Society of Mechanical Engineers, New York, N. Y., vol. 74, no. 6, 1952, pp. 1005–11.

5. Improving the Dynamic Response of Servomechanisms by Second-Derivative Feedback, R. S. Lovett. *S.M.E.E. Thesis*, Massachusetts Institute of Technology, Cambridge, Mass., Aug. 31, 1950.

6. Design of a Single-Stage Hydraulic Servomotor, R. D. Atchley. M.I.T. Dynamic Analysis and Control Laboratory, *Memor Report No. 69*, May 1951. CONFIDENTIAL.

7. Versatility of Floated-Type Rate Integrating Gyroscopes in Systems Applications, J. W. Lower. *Convention Record*, Institute of Radio Engineers, New York, N. Y., pt. 5, 1955.

8. Engineering Applications of Fluid Mechanics (book), J. C. Hunsaker, B. G. Rightmire. McGraw-Hill Book Company, Inc., New York, N. Y., 1947, p. 67.

9. Contributions to Hydraulic Control.—9. Pressure-Flow Relationships for a 4-Way Valve, J. F. Blackburn. *Transactions*, American Society of Mechanical Engineers, New York, N. Y., vol. 75, No. 6, 1953, pp. 1163–70.

10. Reference 9, p. 58.

11. Hydraulic Variable-Speed Transmissions as Servomotors, G. C. Newton, Jr. *Journal*, Franklin Institute, Philadelphia, Pa., vol. 243, no. 6, June 1947, pp. 458, 466.

Discussion

Wayne B. Lloyd (Westinghouse Electric Corporation, Baltimore, Md.): A paper on hydraulic servomechanisms slanted toward the area of hydraulic antenna drives is particularly timely in view of the recent emergence of this type of antenna drive in the field of air-borne radar systems. To one who has been engaged in the use of such equipment it affords a welcome opportunity for comparison of results and conclusions.

The extension of Shearer's analysis to include the effect of mechanical compliance is certainly helpful; however, the omission of viscous damping from the analysis is perhaps unfortunate, because many of these systems have appreciable viscous damping. Possible sources of viscous damping are drag from rubber shaft seals and air drag from large radar reflectors.

The author is correct when he points out that selection of a suitable radian displacement factor is a compromise between various conflicting requirements. The writer's experience indicates that the value of radian displacement factor selected for the experimental system described in the paper, although adequate for good frequency response, leads to a motor size and weight much too large to be incorporated into many types of practical systems. An example of a drive requiring lighter and smaller motors but having comparable load inertia would be a radar antenna gimbaled for 2 degrees of angular freedom.

If size and weight are to be reduced, radian displacement factor must also be reduced, but such a reduction will lower the mechanical resonant frequency unless inertia is also reduced. Since load inertia usually cannot be made substantially lower than that given for the experimental system, one is led to the conclusion that a practical drive of this type will have a low resonant frequency in the valve-motor transfer function.

In many applications closed loop frequency response requirements dictate that the loop crossover frequency must be equal to or greater than the basic mechanical resonant frequency. To accomplish this, some practical and effective means of compensation must be employed. As pointed out by the author, tuned antiresonant electrical networks have limited application because of motor damping variation with temperature. Effective compensation can be achieved by second derivative feedback or by second derivative combined with first derivative feedback. These methods of compensation, however, tend to be costly in terms of size, weight, and complexity when accomplished by use of transducers and electronic components. One type of compensation developed by the writer's company accomplishes a form of acceleration plus velocity feedback through a simple modification of the servo valve.

To demonstrate that practical compensation need not be complex or troublesome, a brief description of this type of valve follows. The valve consists of a spring-centered spool actuated by differential pressure produced by a miniature open-center first stage. The first stage is of the nozzle-flapper variety. The spool is the 3-land type and is constructed so that the center land is smaller in diameter than the two outer lands. This difference in area gives rise to an axial force on the spool proportional to differential pressure across the motor. The direction of this force is opposite in sense to the force produced by the first stage so that the net effect is negative differential pressure feedback. If the load attached to the motor shaft consists of inertia and viscous damping, the motor differential pressure is then proportional to the sum of the acceleration plus the velocity of the load. Therefore, the axial force which is fed back to the spool is a function of the second derivative plus the first derivative of the load angular displacement.

By use of this type of feedback compensation the damping factor of the quadratic introduced by system compliance and inertia can be made larger. Usually it is best to make this resonance term heavily overdamped so that two real roots rather widely separated in frequency result. If this is done, ordinary lead network compensation may be used to achieve closed loop stability. The over-all effect of this type of compensation is somewhat similar to increasing either motor-damping factor or valve-damping factor, or both. However, no variation with temperature occurs, and the total leakage flow rate through the valve at zero stroke is not increased. Moreover, no additional power is consumed when holding against steady-state load torques. Some loss in loop torque constant usually occurs, but closed-loop frequency response and velocity constant are, as a rule, greatly improved by this technique.

It is hoped that this discussion will help to re-emphasize the necessity for devotion of much attention to the mechanical resonance problem inherent in gearless vane-type rotary hydraulic servomotors. This problem, especially important in the case of air-borne equipment where lightness of weight is imperative, presents a real challenge in all applications where inertia loads occur. Also, it is hoped that the paper demonstrates how the solution of this problem need not involve unreasonably troublesome or complicated techniques.

J. L. Shearer (Massachusetts Institute of Technology, Cambridge, Mass.): The author is to be complimented on his work on the development of a high-performance hydraulic control system. For the purpose of discussion, I should like to point out that in many systems of this type it is not possible to neglect the lag in the electromagnetic torque motor which strokes the control valve. This is particularly true in systems where load inertia is relatively small so that lags due to load inertia and fluid compliance are not the major lags in the system.

There seems to be no reason why the author should use Laplace transforms in his exposition of the dynamic characteristics of this system, because he appears to be simply trying to express the differential equations of the system. If his plan was to use these equations for obtaining some sort of transient responses by means of Laplace transform techniques, then, of course, the complex variable s would be needed. However, as far as this paper is concerned, the use of the complex variable s seems superfluous. In his discussion of hydraulic motors the author vividly describes troubles due to mechanical compliance. For the purpose of generality, it would be well to point out also that some motors, particularly piston-type motors, have very large static friction effects and that in some of these motors the leakage is a function of speed as well as of pressure. In fact, the manufacturers' data available at this time on most motors are inadequate for the systems designer who needs to know more quantitatively what performance to expect of a commercially available motor or pump.

It is also important to emphasize the lack of perfect integration of the valve and motor system under conditions of small valve stroke and high load torques, even though the author does discuss his concept of torque constant quite thoroughly. This is particularly important because, although there are some hydraulic motors (i.e., cylinders) that have negligible leakage, it has been impossible, to date, to build a servovalve with no leakage.

K. A. Stone (University of Michigan, Ann Arbor, Mich.): The author's results compare favorably with material to be found in classified reports concerning hydraulic servos published by various aircraft companies.

The amount of damping in the form of friction and motor leakage of high-performance hydraulic servos is normally kept as low as possible for obvious reasons. Motor leakage not only is a source of power loss but it contributes to the positional error when the motor is subjected to load torques. In an attempt to increase the system damping, various valve manufacturers have resorted to underlapped valves. Whether or not this is the main reason this writer does not know. Nevertheless, many valve designers maintain that the valve leakage is an important stabilizing factor in the system. The author has shown that the damping attributable to valve leakage is but a small fraction of the total. Further reflection also indicates that this damping is decreased as the valve moves from its neutral position. In addition, the fact that leakage is present causes the system gain in the underlapped region to be approximately twice that for valve displacements outside the underlapped region. To show this, let us consider equation 34.

$$Q_m = 0.6\frac{\omega}{\sqrt{\rho}}\left[(U+X)\sqrt{P-P_m}-(U-X)\sqrt{P+P_m}\right] \quad |X|<U \quad (71)$$

Then for $X \geqq U$

$$Q_m = 0.6\frac{\omega}{\sqrt{\rho}}\left[(U+X)\sqrt{P-P_m}\right] \quad X \geqq U \quad (72)$$

which for X sufficiently large and $P>>P_m$ becomes equation 35.

$$Q_m = 0.6\omega\sqrt{\frac{P}{\rho}}\,X = k_z X \quad (73)$$

Since the valve gain is taken as the rate of change of Q_m with respect to X, the gain as defined by equation 35 is k_z. As equation 73 (35) is an approximation of equation 2, then

$$\frac{dQ_m}{dX} = 0.6\left\{\frac{\omega}{\sqrt{\rho}}\sqrt{P-P_m}-\frac{1}{2}\frac{U+X}{\sqrt{P-P_m}}\frac{dP_m}{dX}\right\} \quad X \geqq U \quad (74)$$

and if $P>>P_m$ and the term

$$\frac{1}{2}\frac{U+X}{\sqrt{P}}\frac{dP_m}{dX} \approx 0 \quad (75)$$

Then

$$\frac{dQ_m}{dX} \approx 0.6\frac{\omega}{\sqrt{\rho}}\sqrt{P} = k_z \quad (76)$$

Now, considering equation 1 for operation in the underlapped region

$$\frac{dQ_m}{dX} = 0.6\frac{\omega}{\sqrt{\rho}}\left\{\sqrt{P-P_m}+\sqrt{P+P_m}-\frac{1}{2}\left[\frac{U+X}{\sqrt{P-P_m}}+\frac{U-X}{\sqrt{P+P_m}}\right]\frac{dP_m}{dX}\right\} \quad X<U \quad (77)$$

and for $P>>P_m$

$$\frac{dQ_m}{dX} = 0.6\frac{\omega}{\sqrt{\rho}}\left\{2\sqrt{P}-\frac{U}{\sqrt{P}}\frac{dP_m}{dX}\right\} \quad X<U \quad (78)$$

Then, if the second term is compared with equation 5

$$\frac{dQ_m}{dX} \approx \left[0.6\frac{\omega}{\sqrt{\rho}}\sqrt{P}\right]\times 2 \approx 2k_z \quad X<U \quad (79)$$

Thus it is apparent that operation in the underlapped region results in a valve gain or flow sensitivity which is approximately twice that outside of the region, and the general conclusions in the Appendix on valve performance should be modified accordingly.

The fact that the system gain in the underlapped region is approximately twice that for large valve displacements will normally offset any advantages derived from the small amount of damping resulting from valve leakage. The effect of increased gain would not ordinarily appear in an open-loop sinusoidal response test since the valve amplitude is generally greater than the underlap and tends to mask the effect. Also one must realize that these results have been derived for a somewhat idealized case and equation 1 is based on rather severe conditions from the practical point of view. Nevertheless, the effect of increased gain with underlapped valves is present, and it certainly should be considered when optimum performance is essential.

David V. Stallard: The discussers' constructive comments are much appreciated.

As Mr. Lloyd comments, it is somewhat unfortunate that the effect of viscous load damping was neglected. This effect was not important in the motor described in the paper, but it did appear to be important in a newer motor of smaller displacement. It is readily shown that viscous load damping increases the damping ratio in equation 11 without changing the resonant frequency significantly.

Mr. Lloyd's description of his company's special valve for compensating motor resonance should be of interest to many control engineers. This valve and the "pressure-control" valves manufactured by other companies are basically 2-stage flow-control valves with negative feedback of motor pressure to the second stage. It appears possible to analyze the pressure-control valve by a simple extension of the theory in the Appendix; it probably could be represented as in Fig. 3, but with a large value of F_s compared to F_m. Mr. Lloyd's effective technique of compensation avoids the problems of electrical noise and loop phase lag which are encountered in the accelerometer compensation described in reference 6 of the paper. It is hoped that Mr. Lloyd will be able to publish detailed descriptions of some hydraulic systems with pressure feedback.

It is true that some practical and effective means of compensation is usually needed. However, before a control engineer attempts to compensate a hydraulic control system, it is very desirable for him to gain a basic understanding of system dynamics, such as the effect of various compliances on motor-resonant frequency. The exposition and illustration of fundamentals of operation were two objectives of this paper. Furthermore, a basic understanding of motor performance has led to a basic improvement in design. Since the paper was written, the author has developed a new motor with less than half the weight and radian displacement of the large motor of the paper. Because of a considerable reduction in mechanical compliance, the new motor has a slightly higher resonant frequency than the old motor with the same load inertia, despite the inevitable increase in oil compliance.

As Dr. Shearer observes, the lag of the valve torque-motor is appreciable in many systems. The lag in the torque-motor response is caused primarily by two effects: 1. The large inductance of the valve torque-motor makes the current lag behind the amplifier input signal; 2. The mass of the valve piston and friction forces, etc., make the valve stroke lag behind the current and flux changes. The second of these effects was treated in the transfer function from current to stroke. With regard to the first effect, the paper mentions that the valve amplifier was designed with *6AK6* output pentodes of very high plate resistance, so that the lag of valve current behind the amplifier input signal was negligible at 100 cps.

It would be difficult to reply fully to Dr. Shearer's criticism of the use of Laplace transforms without going deeply into mathematical rigor and historical usage. Hence, only a brief reply will be attempted. After deriving the differential equations of the system, the author formed the various transfer functions from them by using the Laplace transformation,[1] so as to facilitate the construction of block diagrams and the analysis of the frequency response. Indeed, the usefulness of block diagrams depends on transfer functions, which are ratios of transformed variables. It would appear to the author that the transform method is theoretically correct and rigorous. Furthermore, it is widely used in the control literature of the AIEE, and it is taught in many of the best modern texts on automatic control.[1] The older method of Leibnitz was to replace $d^n x/dt^n$ by $D^n x$, and similarly, Heaviside's operational method was to replace such a derivative by $p^n x$. The users of these methods then factor out the operator D or p in polynomials, and substitute $j\omega$ to find the sinusoidal frequency response at the frequency ω. Although these older methods have a certain limited usefulness, they do not seem to be rigorous, and do not appear to have much theoretical foundation other than their resemblance to the Fourier and Laplace transforms and the original complex-variable theory of Cauchy. The difference of viewpoint between Dr. Shearer and the author is caused partly by historical usage: The use of Laplace transforms has been promoted chiefly by electrical engineers, such as Gardner[1] and Brown[1]; apparently most mechanical engineers have preferred, as has Dr. Shearer, to use the operator D or p.

Dr. Shearer's point about the friction and leakage characteristics of other motors is well taken. In this paper friction was neglected because the servomotor had a static friction of only about 10 to 15 pound-inches, which did not affect performance appreciably. However, friction in the journal bearings of the newer motor considerably decreased the resonant peak in its frequency response.

As Mr. Stone comments, many commercial values are underlapped. A slight underlap in a 4-way valve is generally preferable to a slight overlap, which causes a deadband. The amount of underlap is usually a compromise between wasteful leakage and expensive manufacturing tolerances. In a well-made valve, the resulting leakage is a very small percentage of the maximum flow rate, and contributes very little damping.

Mr. Stone makes an interesting point about the double flow gain of the valve in the underlapped region. This double gain arises from the fact that the flow in the underlapped region of operation is being controlled by four variable orifices rather than two as in the large-stroke region. The effect has not proved troublesome in the author's research, perhaps because it was masked by valve dithering and/or motor friction.

REFERENCES

1. TRANSIENTS IN LINEAR SYSTEMS (book), M. F. Gardner, J. L. Barnes. John Wiley & Sons, Inc., New York, N. Y., 1942.

2. SERVOMECHANISMS AND REGULATING SYSTEM DESIGN, H. Chestnut, R. W. Mayer. *Ibid.*, vol. 1, 1951; vol. 2, 1955.

3. PRINCIPLES OF SERVOMECHANISMS (book), G. S. Brown, D. P. Campbell. *Ibid.*, 1948.

Synthesis of Feedback Control Systems by Gain-Contour and Root-Contour Methods

VICTOR C. M. YEH
NONMEMBER AIEE

Synopsis: The root-locus method for synthesis of servo systems is modified and extended to cover systems having adjustable time constants. Root-locus and gain-contour methods permit studies of effects upon transient response resulting from variations in gain and variations in time constants to be made independently. The root-contour method facilitates the synthesis of servo systems in which gain is a function of a time constant. Methods for development of gain-contour and root-contour equations for basic servo systems are presented. Feedback charts and illustrative examples are included.

SYNTHESIS of linear feedback control systems in the time domain has depended upon location of closed-loop roots in the complex plane for a given (fixed) set of open-loop roots. Once the closed-loop roots are found, the procedure for obtaining the transient response to a prescribed reference input is straightforward.[1]

The root-locus method introduced by Evans[2] is an expedient method for obing the locus of roots in complex plane when the system gain is being varied. This method has since been further advanced.[3–7] It should be noted, however, for each given configuration of open-loop roots, one root-locus plot must be made, along which the system gain is the one and only adjustable parameter. The flexibility of this method is, therefore, limited in that it cannot easily cope with the type of control problem in which adjustable time constants are essential.

The use of phase-angle loci advanced by Chu[3] has made possible the study of effects of a varying time constant without replotting the entire root locus. This method employs the principle of superposition such that the adjustable element is considered as a modifier which serves to modify the original root locus of the system. However, the final root locus of the system can only be obtained by introducing an intermediate step which involves a considerable graphical construction. As valuable as the method is, nevertheless, it does not compare with the simplicity of the superposition method offered by the frequency analysis.

The methods presented in this paper are as simple and straightforward as the root-locus method. They permit, however, the use of a time constant as a control variable in place of the system gain used in the root-locus method. They do not introduce any intermediate step nor do they require complicated graphical constructions.

The word "locus" shall be associated with systems wherein the open-loop roots are fixed; whereas the word "contour" shall be associated with systems wherein at least one of the open-loop roots is being varied. If the open-loop root varies independently of gain, a gain contour results; if one of the open-loop roots varies as a function of gain, then a root contour results. Single-knob controllers[8] are practical examples which can be analyzed and synthesized by gain and root contours.

Paper 56-211, recommended by the AIEE Feedback Control Systems Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Winter General Meeting, New York, N. Y., January 30–February 3, 1956. Manuscript submitted February 14, 1955; made available for printing November 23, 1955.

VICTOR C. M. YEH is with New York University, New York, N. Y.

The Linear Feedback System

The open-loop transfer function of a linear feedback system having m zeros and n poles may be written as

$$KY(s) = \frac{K \prod_{i=1}^{m} (s+z_i)}{\prod_{i=1}^{n} (s+p_i)} \qquad (1)$$

where

z_i, p_i = zeros and poles of the open-loop transfer function respectively
s = a complex variable

A transfer function having n poles and m zeros shall be referred to as $T(n,m)$; e.g., a 4-pole 1-zero transfer function is designated by $T(4,1)$. The total number of open-loop roots is $h = m+n$.

The closed-loop characteristic equation

$$1 \pm KY(s) = 0 \qquad (2)$$

may be written more explicitly in terms of p_i, z_i, and K as

$$\prod_{i=1}^{n} (s+p_i) \pm K \prod_{i=1}^{m} (s+z_i) = 0 \qquad (3)$$

where the plus and minus signs indicate degenerative and regenerative feedback respectively. Equation 3 may be refactored

$$\prod_{i=1}^{n} (s+q_i) = 0 \text{ for } (n \gtrless m) \qquad (4)$$

where q_i are the roots of the characteristic equation, or the closed-loop poles. When comparing equations 3 and 4, it is seen that closed-loop poles are functions of open-loop roots (poles and zeros) as well as system gain; i.e., in symbols

$$q_i = Q_i(K, p_1, p_2, \ldots p_n, z_1, z_2, \ldots z_m) \qquad (5)$$

The root locus is the locus of q_i when K is chosen as parametric variable. The gain contour shall be defined as the locus

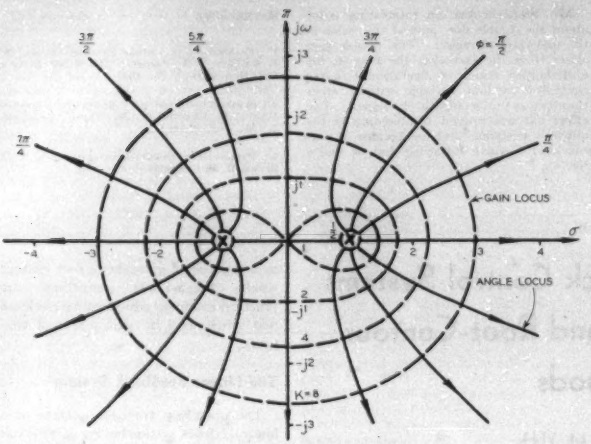


Fig. 1. Gain loci and angle loci of type T(2,0)

of q_i when one and only one of the open-loop roots is chosen as parametric variable and the system gain remains constant. Thus, there are $(m+n)$ sets of gain contours depending on which of the $(m+n)$ open-loop roots is used as parametric variable. The root contour shall be defined as the locus of q_i of a control system in which one of its open-loop roots varies as a function of gain. A more detailed discussion will be presented in the following.

Gain Loci and Angle Loci

The gain loci and angle loci of which root loci are a special case must be distinguished from gain contours and root contours. First, let the co-ordinates of variable and function complex planes be defined as follows:

variable complex plane: $s=\sigma+j\omega$ (6A)

function complex plane: $\dfrac{1}{Y(s)}=Ke^{j\phi}$ (6B)

where K and ϕ are magnitude and phase angle of function complex plane respectively. Constant K and ϕ lines in the s plane are defined as gain loci and angle loci respectively. Fig. 1 shows the conformal mapping of $T(2,0)$ transfer function in the s-plane.[7] It is noted that phase angle varies along a gain locus, whereas gain value varies along an angle locus. Gain loci and angle loci are mutually orthogonal.

Rewriting the characteristic equation in polar form

$$\frac{1}{Y(s)}=Ke^{jl\pi}=\underline{K/l\pi} \qquad (7)$$

or, breaking into angle and magnitude

phase angle: $\underline{/1/Y(s)}=l\pi$ (7A)

magnitude: $\left|1/Y(s)\right|=K$ (7B)

where l is an integer. When comparing equations 6 (B) with 7, it is seen that the particular angle locus with $\phi=l\pi$ satisfies the angle condition of the characteristic equation and is known as the root locus. Since gain is a parameter along root locus the latter completely satisfies the characteristic equation as long as no specific gain value has been assigned. The gain loci, on the other hand, satisfies the characteristic equation only where $\phi=l\pi$, i.e., only at discrete points. It is because of this limitation that gain loci could not be used advantageously in analysis and synthesis of linear feedback control systems. Similarly, root loci could not be used advantageously in cases where adjustable time constants are essential; thus, the development of gain contours and root contours becomes necessary.

Root Locus

It has been shown[7] that root-locus equations may be obtained by considering the angle condition alone. The general expression of the root-locus equation is

$$\sum_{\substack{b=1\\ b=\text{odd}}}^{h'}\left(\sin\frac{b\pi}{2}\right)\sum_{\lambda=1}^{{}_nC_b}(\Pi_b\tan\theta_r)=0 \qquad (8)$$

where

$h'=h$, when h is odd or $=h-1$, when h is even

$$\sum_{\lambda=1}^{{}_nC_b}(\Pi_b\tan\theta_r)$$

is defined as the sum of each combination product.

For simple systems the root loci consist of straight lines, circles, hyperbolas, and various third-degree algebraic curves or combinations thereof. The root-locus equations, as obtained in reference 7, are reproduced in Table I. The notation $L(n,m)$ shall be used to designate root-locus equation of a $T(n,m)$ transfer function.

Gain Contours

Consider a transfer function of the form

$$KY(s)=\frac{1{,}280(S+\alpha)}{(S+2)(S+10)(S+25)} \qquad (9)$$

in which all open-loop poles and gain are given, and α is the only adjustable parameter. Since the system gain has already been specified, the plotting of a complete root locus along which gain varies is of no particular value to the design of this control system. If the roots of the characteristic equation pertaining to the transfer function of equation 9 are plotted in the s plane with α as a parameter, a new contour is obtained along which the system gain is constant. This new contour shall be called a gain contour. Thus, a gain contour is the path or paths traced out by closed-loop roots as they take on new values corresponding to variations in the value of α. The requirements of analytic functions do not permit an arbitrary gain contour; as will be shown, gain contours are algebraic curves.

Suppose in equation 9 that one of the open-loop poles, say p_1, is adjustable instead of α; then a different gain contour in the s plane will be obtained as p_1 takes on new values. To distinguish various types of gain contours, the parameter along a gain contour must be specified. The symbols $G(n,m;\ \alpha_1)$ and $G(n,m;\ p_1)$ shall represent the gain contours, with α_1 and p_1 respectively as the varying parameter, for a transfer function $T(n,m)$.

Gain-Contour Equations

Derivation of a gain contour consists of the following three steps:

1. Obtain the root-locus equation $L(n,\ m)$ for the given system from Table I (or follow

Table I. Root-Locus Equations for $h = m + n \leq 4$

Type	h	Root-Locus Equations (Excluding Real Axis)	Constants in Root-Locus Equations
$L(1,0)$ .. 1 $L(1,1)$ .. 2		none	none
$L(2,0)$ .. 2		line: $\sigma = \frac{1}{2}(\sigma_1 + \sigma_2)$	perpendicular bisector of segment between the open-loop poles
$L(2,1)$ .. 3 $L(2,2)$ .. 4		circle: $(\sigma - \sigma_0)^2 + \omega^2 = R^2$	$\sigma_0 = \frac{\Omega_{12}^2 - B_{12}^2}{(\sigma_1 + \sigma_2) - (\alpha_1 + \alpha_2)}$, $R^2 = \sigma_0^2 + \frac{(\sigma_1 + \sigma_2)B_{12}^2 - (\alpha_1 + \alpha_2)\Omega_{12}^2}{(\sigma_1 + \sigma_2) - (\alpha_1 + \alpha_2)}$
$L(3,0)$ .. 5		hyperbola: $\left(\frac{\sigma - \sigma^*}{a}\right)^2 - \left(\frac{\omega}{b}\right)^2 = 1$	$a = \frac{1}{3}\sqrt{(\sigma_1+\sigma_2+\sigma_3)^2 - 3(\sigma_1\sigma_2 + \sigma_1\sigma_3 + \sigma_2\sigma_3 + \omega_3^2)}$ $b = \sqrt{3}a \quad c = \sqrt{a^2 + b^2} = 2a$
$L(3,1)$ .. 4		third-degree curves: $\omega^2 = (-1)^n \frac{\sigma^3 - a_2\sigma^2 + a_1\sigma - a_0}{\sigma - \sigma_m}$	$a_2 = \frac{1}{2}[(\sigma_1+\sigma_2)\alpha_1\sigma_1 + (\alpha_1 - \sigma_3)\Omega_{12}^2]$ $a_1 = \alpha_1(\sigma_1\sigma_2 + \sigma_1\sigma_3)$ $a_0 = \frac{1}{2}(\sigma_1\sigma_2\sigma_3 + 3\alpha_1)$
$L(4,0)$ .. 4			$a_3 = \frac{1}{4}[(\sigma_1+\sigma_2)\Omega_{12}^2 + (\sigma_3+\sigma_4)\Omega_{34}^2]$ $a_2 = \frac{1}{2}[(\sigma_1+\sigma_2)(\sigma_3+\sigma_4) + \Omega_{12}^2 + \Omega_{34}^2]$ $a_1 = \frac{1}{2}(\sigma_1+\sigma_2+\sigma_3+\sigma_4) = 3\sigma^*$

Centroid	Asymptotic Center	Open-Loop Roots	Natural Frequency
$\sigma^* = \frac{1}{n}\sum_{i=1}^{n} p_i(\sigma_m = \frac{1}{(n-m)}\left(\sum_{i=1}^{n} p_i - \sum_{i=1}^{m} z_i\right)$	...	$p_1 = \sigma_1 + j\omega_1,\ p_2 = \sigma_2 - j\omega_1$	$\Omega_{12}^2 = \sigma_1\sigma_2 + \omega_1^2$
		$z_1 = \alpha_1 + j\beta_{12},\ z_2 = \alpha_1 - j\beta_{12}$	$B_{12}^2 = \alpha_1\alpha_2 + \beta_{12}^2$

Table II. Gain-Contour Equations for $h = m + n \leq 4$

Type	Gain-Contour Equations (Excluding Real Axis)	Explicit Expressions for Closed-Loop Poles q_i
$G(1,0;\sigma_1)$ .. 1 $G(1,1;\sigma_1)$ $G(1,1;\alpha_1)$	none	$\sigma = \sigma_1 \pm K$ none $\sigma = \frac{\sigma_1 \pm K\alpha_1}{1 \pm K}$
$G(2,0;\sigma_m)$ .. 3	horizontal lines: $\omega = \pm\sqrt{\sigma_m^2 - \Omega_{12}^2 \mp K}$	$q_i = \sigma_m \pm \sqrt{\sigma_m^2 - \Omega_{12}^2 \mp K}$
$G(2,1;\alpha_1)$ .. 4	vertical lines: $\sigma = \sigma_m \pm \left(\frac{K}{2}\right)$	$q_i = \left(\sigma_m \pm \frac{K}{2}\right) \pm \sqrt{\sigma_m^2 - \Omega_{12}^2 + \left(\frac{K}{2}\right)[(\sigma_1 - \alpha_1) + (\sigma_2 - \alpha_1)] + \left(\frac{K}{2}\right)^2}$
$G(2,2;\alpha_m)$ .. 5	vertical lines: $\sigma = \frac{\sigma_m \pm K\alpha_m}{1 \pm K}$	$q_i = \frac{(\sigma_m \pm K\alpha_m) \pm \sqrt{(\sigma_m \pm K\alpha_m)^2 - \Omega_{12}^2 \pm K B_{12}^2}}{1 \pm K}$
$G(3,0;\sigma_1)$ .. 6	third-degree curve: $\omega^2 = \frac{1}{2}\Big\{ -[(\sigma-\sigma_1)^2 + (\sigma-\sigma_2)^2 - 2\omega_3^2] \pm \sqrt{[(\sigma-\sigma_1)^2 - (\sigma-\sigma_2)^2]^2 - 8\omega_3^2[(\sigma-\sigma_1)^2 + (\sigma-\sigma_2)^2] \mp K4[(\sigma-\sigma_1) + (\sigma-\sigma_2)]} \Big\}$	
$G(3,1;\alpha_1)$ .. 5	hyperbola: $\left(\frac{\sigma - \sigma^*}{a}\right)^2 - \left(\frac{\omega}{b}\right)^2 = 1$	$\sigma^* = \frac{1}{n}\sum_{i=1}^{n} p_i = \frac{1}{2}(\sigma_1 + \sigma_2 + \sigma_3)$
	$a = \frac{1}{3}\sqrt{(\sigma_1+\sigma_2+\sigma_3)^2 - 3(\sigma_1\sigma_2 + \sigma_1\sigma_3 + \sigma_2\sigma_3 + \omega_3^2) \mp K}$	$b = \sqrt{3}a$

Sign of Feedback	Mean Values	Open-Loop Roots	Natural Frequency
Upper sign in front of K for degenerative	$\sigma_m = \frac{1}{2}(\sigma_1 + \sigma_2)$	$p_1 = \sigma_1 + j\omega_1,\ p_2 = \sigma_1 - j\omega_1$	$\Omega_{12}^2 = \sigma_1\sigma_2 + \omega_1^2$
Lower sign in front of K for regenerative	$\alpha_m = \frac{1}{2}(\alpha_1 + \alpha_2)$	$z_1 = \alpha_1 + j\beta_{12},\ z_2 = \alpha_1 - j\beta_{12}$	$B_{12} = \alpha_1\alpha_2 + \beta_{12}^2$

the procedures which are outlined in reference 7).

2. Select one of the open-loop roots, say p_1, as parameter and solve for it in terms of σ, ω, and other open-loop roots from the root-locus equation $L(n, m)$.

3. Substitute this expression into the magnitude criterion, equation 7(B), and simplify to obtain gain-contour equation $G(n, m; p_1)$.

It is noted that the parameter p_1 is thus eliminated by substitution and that, for each value of gain other than zero, an algebraic curve —the gain contour—in the s plane is obtained. Different points on this countour represent different locations of closed-loop roots as the parameter p_1 takes on new values. The gain value at stability limit or at "aperiodic" limit can be obtained from gain-contour equations by setting $s = 0$; this information could not be obtained from the root-locus equation since K does not appear in that equation. Gain-contour equations for systems having up to four open-loop poles are given in Table II. Examples for the derivation of gain-contour equations may be found in Appendix I.

Illustrative Examples

The significance of gain contours is demonstrated by the following examples. Consider a feedback control system having a $T(2,1)$ transfer function of the form

$$KY(s) = \frac{K(s + \alpha_d)}{(s-1)(s+3)} \tag{10}$$

where both K and the derivative time constant $\tau_d = 1/\alpha_d$ are adjustable. It is noted that the open-loop system is unstable because of a pole in the right half of the s plane. From Tables I and II, the root-locus and gain-contour equations for $T(2,1)$ transfer function are obtained as follows:

The root-locus equation $L(2,1)$ is

$$(\sigma - \alpha_d)^2 + \omega^2 = (\sigma_1 - \alpha_d)(\sigma_2 - \alpha_d) \tag{10A}$$

The gain-contour equation $G(n, m; \alpha_d)$ is

$$\sigma = \left(\frac{\sigma_1 + \sigma_2}{2}\right) + \left(\frac{K}{2}\right) \tag{10B}$$

which shows that root loci are circles and gain contours are vertical lines. Fig. 2 shows two root loci (solid lines) and several gain contours (dash lines) of $T(2,1)$. Suppose that, originally, the derivative time constant is set at $\tau_d = 1/\alpha_d = 0.20$ second; then, upon the closure of feedback loop, the two-poles will first move toward each other along the real axis and meet at point E and then split and move

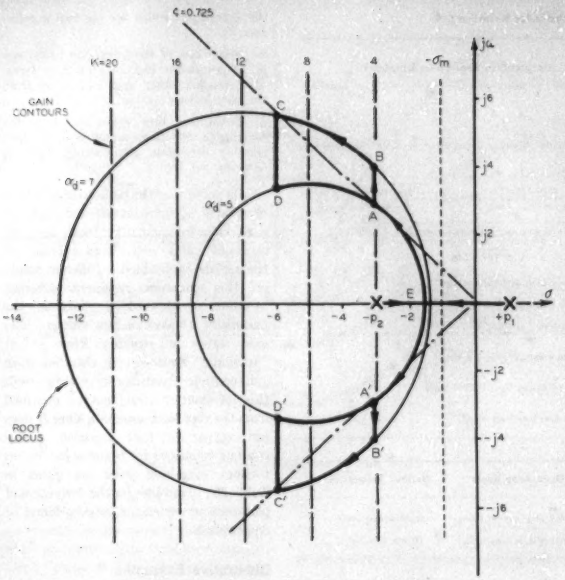


Fig. 2. Root loci and gain contours of type T(2,1)

along the root-locus circle $\alpha_d=5$ as gain is increased. When the pole p_1 (which later becomes q_1 when the loop is closed) moves across the imaginary axis, the system becomes stable; this occurs at $K=3/5$. For a specified damping ratio $\zeta=0.725$, a value of gain $K=4$ is required. Under the foregoing conditions, $\alpha_d=5$ and $K=4$. The closed-loop poles are located at A and A'; see Fig. 2. Since the $\zeta=0.725$ line is tangent to the circle $\alpha_d=5$ at points A and A', further increase in gain will cause an increase in damping ratio. Hence, at this point, the speed of response cannot be improved by adjustment of gain. An increase in α_d from 5 to 7, however, will enable the poles to move vertically from A, A' to B, B' (as shown by arrows) thus decreasing the damping ratio and improving the speed of response. If, however, the original damping ratio $\zeta=0.725$ is to be retained, a further increase in gain from 4 to 10 could be made along the new root-locus circle $\alpha_d=7$, so that the final pole locations are at C and C'. An alternative way to improve the system is to increase the gain from zero to 10 (poles at D and D') and then change α_d from 5 to 7 (final poles at C and C'). It is also noted that the difference in gain values between BC and AD are the same since AB and CD are part of gain contours. Similarly, the increase in gain value from open-loop conditions to A and that to B are the same.

Speed-Control System

To illustrate further the use of root locus and gain contours, a speed-control system, Fig. 3, is now being considered. Here, $c(t)$ and $r(t)$ are variations of controlled speed and desired speed, and T_m and T_L are motor torque and load torque respectively. The output member having inertia J is driven by a 2-phase motor, the transfer function of which is approximated by a simple lag and a gain factor k_m. A proportional-plus-integral controller is used, the transfer-function of which is

$$G(s)=k\left(1+\frac{1}{\tau_i s}\right)=\frac{k(s+\alpha_i)}{s} \qquad (11)$$

where $\tau_i=1/\alpha_i$ is the integral time constant. The open-loop transfer function is then

$$KY(s)=\frac{K(s+\alpha_i)}{s^2(s+\sigma_m)} \qquad (12)$$

where $K=k(k_m/J)=10k$. The transformed response function is

$$c(s)=\frac{K(s+\alpha_i)r(s)-\sigma(s+\sigma_m)T_L(s)}{(s+q_1)(s+q_2)(s+q_3)} \qquad (13)$$

where q_1, q_2, and q_3 are the roots of the characteristic equation.

The root loci and gain contours of this system are shown in Fig. 4. It can be shown that the zero $(-\alpha_i)$ must have values between 0 and -12 or the closed-loop system will be unstable. When the zero is at -12, it cancels the pole contributed by the motor, and the root locus of the system coincides with the $j\omega$ axis. Thus, oscillation prevails regardless of system gain. As τ_i is increased, the zero moves toward the origin and the root locus bends to the left. When $\tau_i=3/4$ second, the locus becomes a strophoid. Further increase in τ_i will cause the locus to split into two parts. Finally, when τ_i approaches infinity, a simple proportional controller results and the locus consists of the vertical line AA' and the real axis. The gain contours for $T(3,1)$ are hyperbolas as shown, and degenerate to straight lines at $K=48$ for the particular example.

The transient response to a step function change in reference input is shown in Fig. 5 for the following three conditions:

Case	K	τ_i Seconds	α_i	Roots of Characteristic Eq
a	48	1	−1	−1.48, −5.23±*j*2.20
b	48	3/4	−4/3	−4, −4, −4
c	48	1/2	−2	−7.20, −2.40±*j*2.75

From Fig. 5, the following transient response data may be obtained:

Case	T_b, Seconds	T_p, Seconds	T_s, Seconds	Per-Cent Overshoot	Per-Cent Undershoot	Control Area
a	0.08	0.33	2.5	71	none	−
b	0.21	0.75	2.0	25	none	0
c	0.76	1.20	1.7	17	34	+

where

T_b = build-up time = time when response first crosses the unity line

T_s = settling time = time for response to reach within 2 per cent of final value

T_p = peak time = time at the maximum transient peak

From the foregoing results it is seen that case *b* is the most desired control condition. Both the speed of response and per cent of overshoot are adequate, and there will be no steady-state error. The positive and negative parts of the control area (shaded) cancel each other, leaving a net total of zero control area.

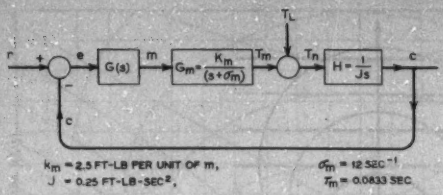


Fig. 3 (above) A speed-control system

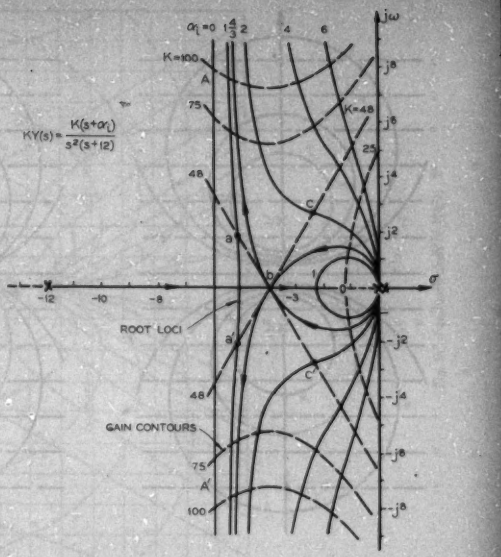


Fig. 4 (right). Root loci and gain contours of type T(3,1)

It is to be noted that the transient response, in this example, is extremely sensitive to the value of the integral time constant. A slight deviation from the optimum value will result in an unacceptable transient response, as shown by curves *a* and *c* in Fig. 5; this illustrates the delicate nature of the control.

The Feedback Chart

Consider the following two transfer functions:

For $T(3,0)$

$$KY(s) = \frac{K}{(s+\sigma_1+j\omega_{12})(s+\sigma_2-j\omega_{12})(s+\sigma_3)} \tag{14}$$

For $T(3,1)$

$$KY(s) = \frac{K(s+\alpha)}{(s+\sigma_1+j\omega_{12})(s+\sigma_2-j\omega_{12})(s+\sigma_3)} \tag{15}$$

in which either $\omega_{12}=0$, which gives three real roots, or $\sigma_1=\sigma_2$, which gives one real and two complex roots. Both the root-locus equation of $T(3,0)$ and the gain-contour equation of $T(3,1)$ are identical in form, see Tables I and II, and are given by

$$\omega^2 = 3[(\sigma-\sigma^*)^2 - a^2] \tag{16}$$

where the centroid $\sigma^* = (1/n)\left(\sum_{i=1}^{n} p_i\right)$ and the value of a are given by the following expressions:

For $L(3,0)$

$$a = \frac{1}{3}\times \sqrt{(\sigma_1+\sigma_2+\sigma_3)^2 - 3(\sigma_1\sigma_2+\sigma_2\sigma_3+\sigma_3\sigma_1+\omega_{12}^2)} \tag{17A}$$

For $G(3,1;\ \alpha)$

$$a = \frac{1}{3}\sqrt{(\sigma_1+\sigma_2+\sigma_3)^2 - 3(\sigma_1\sigma_2+\sigma_2\sigma_3+\sigma_3\sigma_1+\omega_{12}^2) \mp 3K} \tag{17B}$$

Fig. 6 is a chart showing a family of hyperbolas given by equation 16 as a takes on different values. It may be considered as a root-locus chart if equation 17(A) is used, or as a gain-contour chart if equation 17(B) is used. In Fig. 6, the centroid σ^* is taken as the center; the $j\omega$ axis may be located relative to the centroid, thus making the chart universally applicable. By calculating a for $T(3,0)$ or $T(3,1)$, its root locus, or gain contours, may be obtained immediately from Fig. 6.

A feedback chart for a given transfer function is obtained if the root-locus chart is superposed upon the gain-contour chart of the same transfer function. A typical example has already been shown in Fig. 4, where the analysis of a speed-control system was considered. A numerical example will be found in Appendix II, where the use of universal feedback charts is shown.

For the $T(2,1)$ and $T(2,2)$ feedback charts, the root loci are circles, and the gain contours are vertical lines. Fig. 7

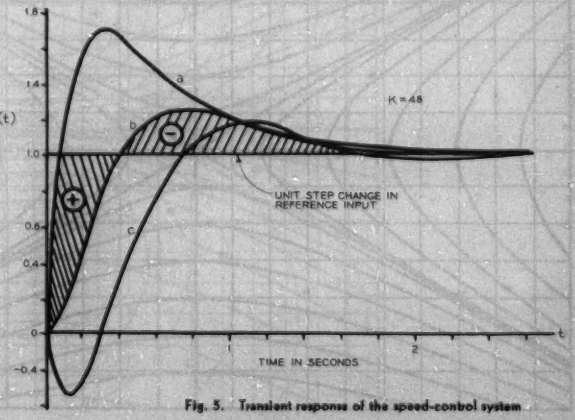


Fig. 5. Transient response of the speed-control system

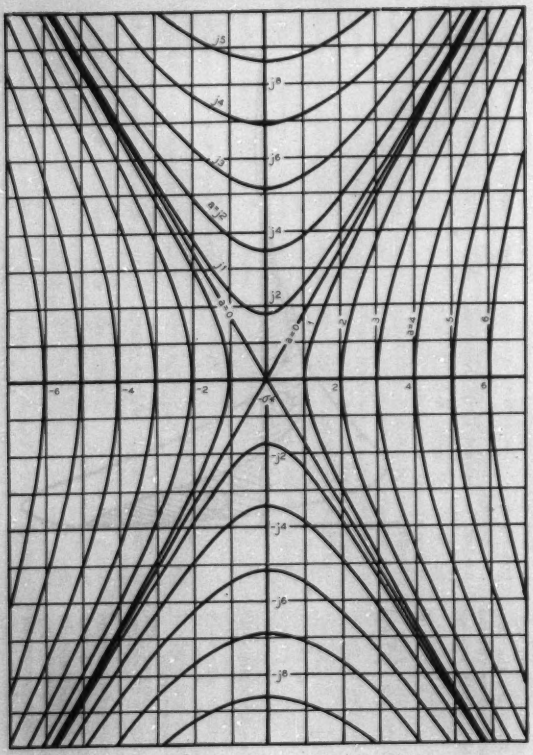


Fig. 6. Root-locus chart L(3,0) or gain-contour chart G(3,1)

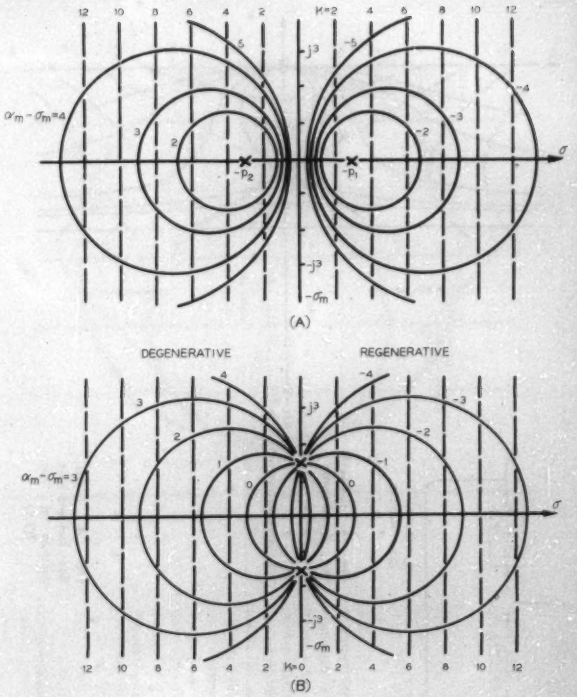


Fig. 7. Root loci and gain contours of type T(2,2)

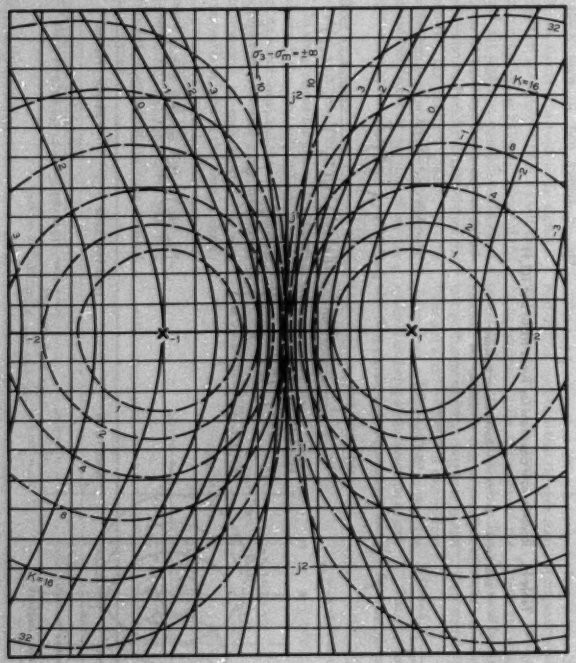


Fig. 8. Feedback chart of T(3,0)

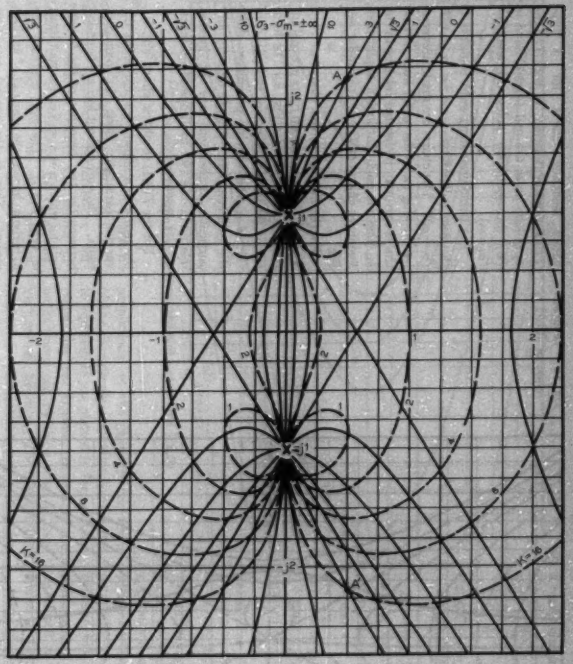


Fig. 9. Feedback chart of T(3,0)

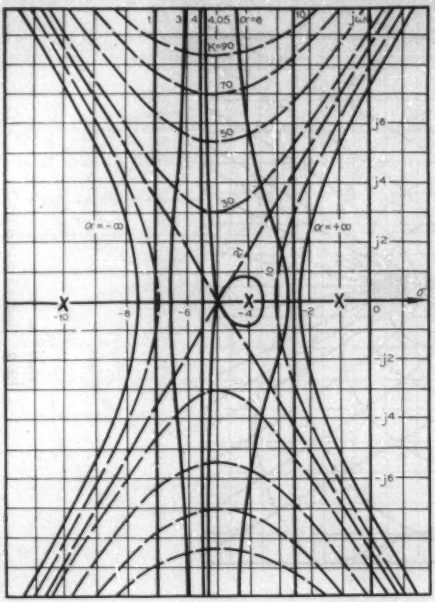


Fig. 10. Feedback chart of T(3,1)

shows feedback charts for a type $T(2,2)$ transfer function; in Fig. 7(A) the open-loop poles are real ($\omega_{12}=0$), and in (B) the open-loop poles are complex ($\sigma_1=\sigma_2$).

The parameter along gain contours is the mean value of two zeros, i.e., $\alpha_m=1/2-(\alpha_1+\alpha_2)$. The charts are made universal by choosing σ_m as the center lines, so that the $j\omega$ axis may be located relative to the σ_m line for each particular application. Feedback charts of $T(2,1)$ are similar, see Fig. 2.

For the $T(3,0)$ feedback charts see Figs. 8 and 9. The open-loop roots are real for Fig. 8 but complex for Fig. 9. The location of the third pole σ_3, not shown is considered as the varying parameter along the gain contours. The $j\omega$ axis for any particular application may be located relative to the two fixed open-loop poles. See example in Appendix II.

For the $T(3,1)$ feedback chart, Fig. 10, the zero α (not shown) was chosen as the varying parameter along the gain contours. The various root loci are bounded between the two branches of the hyperbola for which $\alpha=\pm\infty$, i.e., the root locus of type $T(3,0)$. The gain contours are hyperbolas having the centroid σ^* as their centers. This is the special condition where the root locus becomes a strophoid and the gain contour becomes two straight lines all passing through the centroid. Fig. 10 applies only to cases where

Table III. Root-Contour Equations for Types T(2,0) and T(2,1)

Case	Transfer Function	Root Locus (Excluding $\omega=0$)	Root Contour Equation (Excluding $\omega=0$)	Explicit Expressions for Closed-Loop Poles q_i
			parabola:	
A	$\dfrac{K}{(s+\sigma_1+K+j\omega_{12})(s+\sigma_2+K-j\omega_{12})}$	$\sigma=\sigma_m\pm K$	$\sigma=\omega^2+(\sigma_m-\Omega_{12}^2)$	$q_i=(\sigma_m\pm K)\pm\sqrt{\sigma_m^2-\Omega_{12}^2\mp K}$
			circle:	
B	$\dfrac{K}{(s+\sigma_1)(s+K)}$	$\sigma=\dfrac{\sigma_1+K}{2}$	$[\sigma-(\sigma+1)]^2+\omega^2=(\sigma+1)$	$q_i=\dfrac{\sigma_1\pm K}{2}\pm\sqrt{\left(\dfrac{\sigma_1-K}{2}\right)^2\mp K}$
			third-degree curve:	
C	$\dfrac{K}{(s+\sigma_1)\left(s+\sigma_2+\dfrac{1}{K}\right)}$	$\sigma=\sigma_m\pm\dfrac{1}{2K}$	$\omega^2=\dfrac{1}{[(\sigma-\sigma_1)+(\sigma-\sigma_2)]}-(\sigma-\sigma_1)^2$	$q_i=\left(\sigma_m+\dfrac{1}{2K}\right)\pm\sqrt{\left(\dfrac{\sigma_1-\sigma_2}{2}\right)^2+\dfrac{\sigma_m-\sigma_1}{K}+\left(\dfrac{1}{2K}\right)^2\mp 1}$
			hyperbola:	
D	$\dfrac{K(s+K)}{(s+\sigma_1+j\omega_{12})(s+\sigma_1-j\omega_{12})}$	$(\sigma-K)^2+\omega^2=(\sigma_1-K)(\sigma_1-K)+\omega_{12}^2$	$\left\{\begin{array}{l}\left(\sigma-4/3\sigma_m\right)^2-\left(\dfrac{\omega}{b}\right)^2=1\\ a=\dfrac{1}{3}\sqrt{(\sigma_1+\sigma_2)^2-3\Omega_{12}^2},\ b=\sqrt{3}a\end{array}\right\}$	$q_i=\left(\sigma_m\pm\dfrac{K}{2}\right)\pm\sqrt{\sigma_m^2-\Omega_{12}^2\pm K(\sigma_m+K)+\left(\dfrac{K}{2}\right)^2}$,
			circle:	
E	$\dfrac{K\left(s+\dfrac{1}{K}\right)}{(s+\sigma_1+j\omega_{12})(s+\sigma_1-j\omega_{12})}$	$\left(\sigma-\dfrac{1}{K}\right)^2+\omega^2=\left(\sigma_1-\dfrac{1}{K}\right)\left(\sigma_2-\dfrac{1}{K}\right)+\omega_{12}^2$	$\sigma^2+\omega^2=\Omega_{12}^2$	$q_i=\left(\sigma_m\pm\dfrac{1}{2K}\right)\pm\sqrt{\sigma_m^2-\Omega_{12}^2\pm K(\sigma_m-K)+\left(\dfrac{1}{2K}\right)^2}$

Signs in the Feedback Path	Mean Values	Open-Loop Roots	Natural Frequency
Upper sign in front of K for degenerative feedback	$\sigma_m=\dfrac{1}{2}(\sigma_1+\sigma_2)$	$p_1=\sigma_1+j\omega_{12},\ p_2=\sigma_2-j\omega_{12}$	$\Omega_{12}^2=\sigma_1\sigma_2+\omega_{12}^2$
Lower sign in front of K for regenerative feedback	$\alpha_m=\dfrac{1}{2}(\alpha_1+\alpha_2)$	$z_1=\alpha_1+j\beta_{12},\ z_2=\alpha_2-j\beta_{12}$	

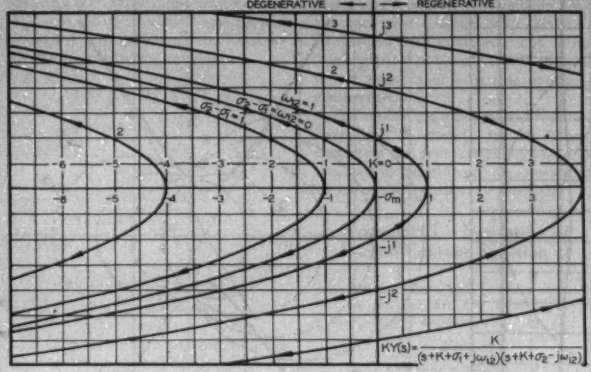


Fig. 11. Root contours, gain contours, and root loci of type T(2,0)

$(p_3-p_2)/(p_3-p_1)=2$; hence, it is not a universal feedback chart for all $T(3,1)$ transfer functions. Nevertheless, all $T(3,1)$ feedback charts are similar in form; this is evidenced by comparing Figs. 4 and 10. For more complex systems, feedback charts are no longer universal, and individual plots must be made for each application.

The cubic charts and quartic charts developed by Liu[9] and Evans[10] differ from the feedback charts presented here. In the former, the roots of an algebraic equation are obtained from a chart used in conjunction with the coefficients of the relevant algebraic equation. The effect upon the roots of varying any of the coefficients could not be visualized from the charts. The feedback charts, on the other hand, show clearly the effects on closed-loop roots of varying either the gain or the open-loop roots.

Root Contours

In the previous discussion, it has been assumed that the adjustment of gain along a root locus and the adjustment of a time constant along a gain contour are independent of each other. The root contour is defined as the locus of closed-loop roots of systems in which the adjustment of gain and the time constant are interrelated; i.e., the system gain is a function of an adjustable time constant or vice versa.

Consider the transfer function:

$$KY(s)=\frac{K(s+K)}{(s+\sigma_1+j\omega_{12})(s+\sigma_2-j\omega_{12})(s+\sigma_3)} \qquad (18)$$

in which the location of the zero $\alpha(=K)$ is proportional to the gain. If the roots of the characteristic equation pertaining to equation 18 are plotted in the complex plane, the root contour consists of the paths traced out by these roots as they vary with K, hence α. The root-contour method is particularly well suited for analysis and synthesis of pneumatic controllers in which an adjustable time constant is often a function of gain.[11] Single-knob controllers[8] are examples.

The root-locus method differs from the root-contour method in that the open-loop roots considered in the former are all fixed; whereas, in the latter, one of the open-loop roots varies with gain. The methods are similar in that system gain is a parameter varying along both a root locus and a root contour. A root contour differs from a gain contour in that the system gain being a function of an adjustable time constant, is not constant along a root contour, while it remains constant along a gain contour. They are similar in that both have an open-loop root as a varying parameter.

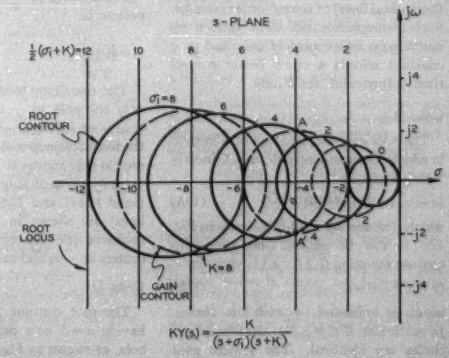


Fig. 12. Root contours and root loci of type T(2,0)

Root-Contour Equations and Basic Forms

The steps in the derivation of root-contour equations are outlined as:

1. By following procedures outlined in reference 7, obtain root-locus equation $L(n, m; \alpha_i)$ or $L(n, m; p_i)$, wherein either α_i or p_i is a function of gain K.
2. Solve root-locus equation for K in terms of σ, ω, and other open-loop roots.
3. Substitute into equation 7(B), eliminating K. The result, when simplified, is the root-contour equation $R(n, m; \alpha_i)$ or $R(n, m; p_i)$.

The root-locus equation obtained in step 1 contains the system gain K. For a given value of gain, there are only n points on the locus satisfying the characteristic equation. Just as the gain contour was developed to replace the gain locus, the root contour must be developed to replace the root locus, as shown in steps 2 and 3.

Five different types of transfer functions are considered here, all of which are second-order systems. Their root-locus and root-contour equations are tabulated in Table III. Since they are used only to illustrate the principle involved, no complex systems are treated. Moreover, in the examples chosen, the open-loop roots are considered either linear or reciprocal functions of gain. There are infinite varieties of functional relationship; their derivations are straightforward.

Case A

The root-locus and root-contour equations are given in Table III. In addition

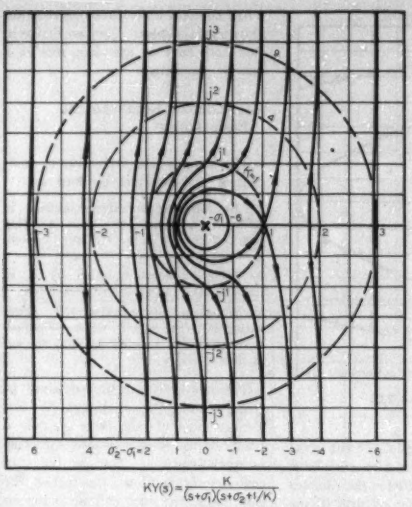


Fig. 13. Root contours and gain contours of type T(2,0)

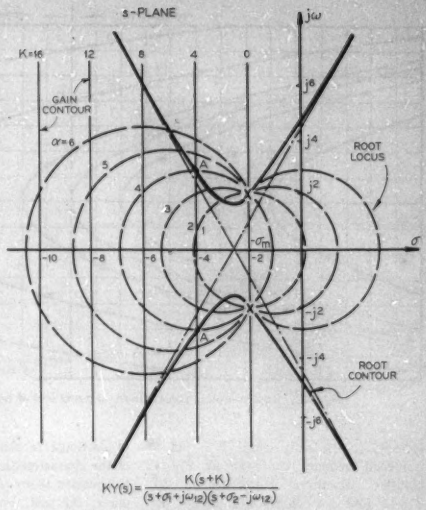


Fig. 14. Root contours and root loci of type T(2,0)

to the real axis, root loci consist of vertical lines, and root contours are parabolas, as shown in Fig. 11. For an initially critically damped system ($\sigma_1=\sigma_2$, and $\omega_{12}=0$) the parabola is tangent to the σ_m line. Increasing ω_{12} causes a rightward shift of the parabola, whereas increasing $1/2(\sigma_1-\sigma_2)$ causes a leftward shift. Degenerative feedback will cause the poles to move to the left along parabolas, whereas regenerative feedback will cause them to move to the right, as indicated by arrows. The root contours of this system should be compared with the root loci (vertical lines) and gain contours (horizontal lines) of second-order systems. Such a comparison will show that a simultaneous adjustment of gain and time constant affords a much better control than adjustment separately.

Case B

The root-contour equation of a system in which one open-loop pole varies linearly with gain is

$$[\sigma-(\sigma_1+1)]^2+\omega^2=(\sigma_1+1) \tag{19A}$$

which is shown by solid-line circles in Fig. 12. It can be shown that the gain-contour equation $G(2,0;\ \sigma_1)$ is given by

$$(\sigma-K)^2+\omega^2=K \tag{19B}$$

which is indicated by dash-line circles. In particular, if $K=\sigma_1+1$, the two sets of circles are identical. The system gain does not appear in equation 19(A); hence it is a parameter along the root contour. The open-loop pole σ_1 does not appear in, 19(B); hence it is a parameter along the gain contour. Also, from Table III, the root-locus equation for case B is $\sigma=1/2$ (σ_1+K) wherein both σ_1 and K are present. Points A and A' are located on the root contour for which $\sigma_1=4$. The gain associated with these roots may be determined from either a root locus or a gain contour.

Case C

From Table III, the second pole of the system is

$$p_1=\sigma_2+\frac{1}{K} \tag{20}$$

The open-loop system $K=0$ possesses only one pole at $-\sigma_1$ while the second pole is at infinity. After the closure of the feedback loop and increase of gain, the second pole moves to the right as an inverse function of gain. The root contour (solid lines) and the root loci (vertical lines) are shown in Fig. 13. The gain contours are concentric circles having centers at $-\sigma_1$ and radii of $\sqrt{K}$.

Case D

The root contour of a $T(2,1)$ system having $\alpha=K$ as a parameter is a hyperbola, as shown in Fig. 14. The root loci are circles passing through the two open-loop poles and the gain contours are vertical lines. If the value of α is fixed, the root locus becomes a circle with α as its center[7]; this should be compared with the hyperbola when α is allowed to vary directly with the gain. There is a remarkable difference between the closed-loop poles in the two cases.

Case E

The root contours are concentric circles (not shown). There are an infinite variety of root contours. The foregoing examples suffice for illustration.

Interchangeability

It was pointed out that both the gain contours of $T(3,1)$ and the root loci of $T(3,0)$ are hyperbolas, as shown in Fig. 6; hence only one chart will be sufficient.

As an example, consider a third-degree algebraic equation of the form

$$s^3+a_2s^2+a_1s+a_0=0 \tag{21}$$

This equation may be written:

$$s(s^2+a_2s+a_1)+a_0=0 \tag{22A}$$

$$s^2(s+a_2)+a_1\left(s+\frac{a_0}{a_1}\right)=0 \tag{22B}$$

$$s^2+a_2\left(s^2+\frac{a_1}{a_2}s+\frac{a_0}{a_2}\right)=0 \tag{22C}$$

With reference to equations 2 and 3, the foregoing set of equations may be considered as the characteristic equations of a feedback system having the following transfer functions respectively:
For $T(3,0)$

$$KY(s)=\frac{K}{s(s^3+a_2s+a_1)},\quad K=a_0 \tag{23A}$$

For $T(3,1)$

$$KY(s)=\frac{K\left(s+\frac{a_0}{K}\right)}{s^2(s+a_2)},\quad K=a_1 \tag{23B}$$

For $T(3,2)$

$$KY(s)=\frac{K\left(s^3+\frac{a_1s}{K}+\frac{a_0}{K}\right)}{s^3},\quad K=a_2 \tag{23C}$$

Consider a_0 to be a parameter of the characteristic equation, then:

1. From equation 23(A) with $K=a_0$, a root locus of $T(3,0)$ is generated by varying a_0.

2. From equation 23(B) with $K=a_1$, a gain contour is generated with an adjustable zero at $-a_0/a_1$.

3. From equation 23(C) with $K=a_2$, a second gain contour is generated with two adjustable zeros the product of which varies as a_0/K, and the sum is a_1/K.

Since all three transfer functions are derived from equation 21, the aforementioned three different types of loci in the s-plane with a_0 as the parameter must be the same. Hence, for the foregoing transfer functions having a_0 as parameter, the root locus $L(3.0)$, and the gain contours $G(3,1;\ \alpha)$ and $G(3,2;\ \alpha_1,\ \alpha_2)$ are all identical. The same is true of a_1 and a_2. Hence, the following identities:

Parameter	T(4, 0)	T(3, 1)	T(3, 2)
a_0	root locus	gain contour	gain contour
a_1	gain contour	root contour	gain contour
a_2	gain contour	gain contour	root contour

These relationships will be found useful in synthesis and in the determination of roots of algebraic equations and can be extended to higher order systems.

Conclusion

The gain-contour method complements the root-locus method in analysis and synthesis of control systems. The former applies to systems with adjustable time constants and the latter to systems wherein the gain and time constants are varied simultaneously by a single adjustment. Feedback charts show the influences of two independent variables upon the closed-loop roots. Different feedback charts may be obtained by superposing root loci upon gain contours, root loci upon root contours, and gain contours upon root contours. The terms "generalized root-locus method" may be used to include root-locus, gain, and root-contour methods.

The equations for generalized root loci have been developed. In the future, computers and function plotters used in conjunction with these equations will greatly assist servo system designers.

Appendix I. Derivation of Gain Contour Equation

As an example, a gain contour equation will be derived for the following type $T(3,1)$ transfer function

$$KY(s)=\frac{K(s+\alpha)}{(s+\sigma_1+j\omega_{13})(s+\sigma_2-j\omega_{13})(s+\sigma_3)} \tag{24}$$

Here the transfer function possesses four open-loop roots. Using the general expression of the root-locus equation 8, with $h=4$, the angle condition becomes

$$\tan\theta_1+\tan\theta_2+\tan\theta_3+\tan\theta_4-\tan\theta_1\tan\theta_2\tan\theta_3-\tan\theta_1\tan\theta_2\tan\theta_4-\tan\theta_2\tan\theta_3\tan\theta_4-\tan\theta_1\tan\theta_3\tan\theta_4=0 \tag{25}$$

where

$$\tan\theta_1=\frac{\omega-\omega_{13}}{\sigma_1-\sigma}\qquad \tan\theta_2=\frac{\omega+\omega_{13}}{\sigma_2-\sigma}$$

$$\tan\theta_3=\frac{\omega}{\sigma_3-\sigma}\qquad \tan\theta_4=\frac{-\omega}{\alpha-\sigma}$$

Substituting these expressions into equation 25 and simplifying, the following root-locus equation is obtained

$$\omega\{[2\sigma-(\sigma_1+\sigma_2)][(\sigma-\alpha)(\sigma-\sigma_3)+\omega^2]+(\sigma_3-\alpha)[(\sigma-\sigma_1)(\sigma-\sigma_2)-(\omega^2-\omega_{13}^2)]\}=0 \tag{26}$$

Let α be the parameter along the gain contours. Solve for α from the root-locus equation

$$\sigma-\alpha=-\frac{\{\omega^3[(\sigma-\sigma_1)+(\sigma-\sigma_2)+(\sigma-\sigma_3)]+(\sigma-\sigma_3)\omega_{13}^2+(\sigma-\sigma_1)(\sigma-\sigma_2)(\sigma-\sigma_3)\}}{(\sigma-\sigma_1)(\sigma-\sigma_2)+(\sigma-\sigma_2)(\sigma-\sigma_3)+(\sigma_3-\sigma_1)-(\omega^2-\omega_{13}^2)} \tag{27}$$

The magnitude criterion, equation 7(B), of this system must also be satisfied, i.e.

$$K^2=\frac{[(\sigma-\sigma_1)^2+(\omega-\omega_{13})^2][(\sigma-\sigma_2)^2+(\omega+\omega_{13})^2][(\sigma-\sigma_3)^2+\omega^2]}{[(\sigma-\alpha)^2+\omega^2]} \tag{28}$$

Substitute $(\sigma-\alpha)$ of equation 27 into the equation 28 and simplify

$$(\sigma-\sigma_1)(\sigma-\sigma_2)+(\sigma-\sigma_2)(\sigma-\sigma_3)+(\sigma-\sigma_3)(\sigma-\sigma_1)-(\omega_3-\omega_{13}^2)=\mp 3K$$

Further simplification leads to the gain-contour equation

$$\left(\frac{\sigma-\sigma^0}{a}\right)^2-\left(\frac{\omega}{b}\right)^2=1 \tag{29}$$

where

$$\sigma^0=\frac{1}{3}(\sigma_1+\sigma_2+\sigma_3)$$

$$a=\frac{1}{3}\sqrt{(\sigma_1+\sigma_2+\sigma_3)^2-3(\sigma_1\sigma_2+\sigma_2\sigma_3+\sigma_3\sigma_1+\omega_{13}^2)\mp 3K}$$

$$b=\sqrt{3}a$$

For the gain contours, see Figs. 4 and 10.

Appendix II. Example of Using Feedback Chart

Feedback charts are designed for universal applications; thus, it is in general necessary to normalize a transfer function before using charts. Consider the following:

$$KY(s)=\frac{K}{(s+3+j.1.5)(s+3-j.1.5)(s+7.5)} \tag{30}$$

Normalize the foregoing expression by dividing each factor by $\omega_{13}=1.5$

$$K'Y(s')=\frac{K'}{(s'+2+j.1)(s'+2-j.1)(s'+5)} \tag{31}$$

where $K'=K/\omega_{13}^3$ and $s'=s/\omega_{13}$. The feedback chart for $T(3,0)$, shown in Fig. 9, is now applicable to equation 31. The closed-loop poles may be determined immediately if the gain is given. Since the mean value $\sigma_m'=1/2(\sigma_1'+\sigma_2')=2$, the $j\omega$ axis is located two units to the right from the σ_m line on the chart.

Suppose $K=54$, then $K'=54/(1.5)^3=16$; and $\sigma_3'-\sigma_m'=5-2=3$. From Fig. 9, the intersection of $(\sigma_3-\sigma_m)=3$, and the $K=16$ curve gives two closed-loop poles q_1', $q_2'=-1.50\pm j2.14$ which are indicated by points A and A'. The third pole q_3' may be determined from the relation

$$\sum_{i=1}^{n}p_i'=\sum_{i=1}^{n}q_i'$$

or $q_3'=(2+2+5)-(1.50+1.50)=6.0$

Finally, the true locations of the closed-loop roots are

$q_1=1.5(-1.50+j2.14)=-2.25+j3.21$

$q_2=1.5(-1.50-j2.14)=-2.25-j3.21$

$q_3=1.5(-6.0+j0)=-9.0+j0.$

If σ_3 is made to vary, it is only necessary to follow a new $(\sigma_3-\sigma_m)$ line to determine the closed-loop roots for a given σ_3. Other feedback charts may be used similarly.

References

1. Transient in Linear Systems (book), M. F. Gardner, J. L. Barnes. John Wiley & Sons, Inc., New York, N. Y., 1942

2. Control System Synthesis by Root Locus Method, Walter R. Evans. *AIEE Transactions*, vol. 69, pt. I, 1950, pp. 66–69.

3. Synthesis of Feedback Control System by Phase-Angle Loci, Yaohan Chu. *Ibid.*, vol. 71, pt. II, Nov. 1952, pp. 330–39.

4. Instrument Engineering (book), C. S. Draper, W. McKay, S. Lees. Vol. II. McGraw-Hill Book Company, Inc., New York, N. Y., 1952.

5. Control System Dynamics (book), W. R. Evans. *Ibid.*, 1954.

6. Contribution to the Root-Locus Analysis of the Feedback Control System, Akira Nomoto. *Proceedings*, Second Japan National Congress for Applied Mechanics, Tokyo, Japan, 1952, pp. 359–62.

7. The Study of Transients in Linear Feedback System by Conformal Mapping and Root-Locus Method, V. C. M. Yeh. *A.S.M.E. Transactions*, American Society of Mechanical Engineers, New York, N. Y., vol. 76, 1954, pp. 349–61.

8. Single-Knob Adjustment of Reset Rate and Gain, W. I. Caldwell. *Instruments*, Pittsburgh, Pa., vol. 25, Feb. 1952, p. 196.

9. Servomechanism Charts for Varelying Their Stability and for Finding the Roots of Their Third and Fourth Degree Characteristic Equations, Y. J. Lin. Massachusetts Institute of Technology, Cambridge, Mass., 1941.

10. Solution of Cubic Equations and Cubic Charts, L. W. Evans. *Ibid.*, 1941.

11. Analysis of Pneumatic Controllers, J. M. L. Janssen. Automatic and Manual Control, Butterworths Scientific Publication, Ltd., London, England, 1952, pp. 189–204.

Discussion

C. H. Dawson (University of Rochester, Rochester, N. Y.): As Prof. Yeh and others broaden the application of root locus and related methods to the problem of synthesizing feedback control systems, it is important to try to keep the designer's ultimate goal in mind, i.e., to synthesize a control system which can give the desired performance over the entire range of inputs to which the system may be subjected.

The basic techniques upon which the designer can draw include polar plots of the open-loop transfer function in either the direct or inverse Nyquist planes, Bode presentations of log-magnitude and phase angle against log frequency, and generalized root-locus plots.

The Nyquist plots require extensive computational effort and do not permit ready evaluation of the effects of changing parameters. The gain can be chosen to satisfy a specified closed-loop frequency-response peak which is sometimes closely and sometimes not so closely related to the system transient response. However, the inverse Nyquist plot facilitates consideration of systems with a single inner loop, and both the Nyquist methods seem particularly adapted to describing function treatment of systems having a single nonlinearity.

The Bode method requires a minimum of computation and the effects of changing parameters are always evident. However, nonminimum-phase functions and multiple-loop systems are difficult to treat. In the Bode method, gain is chosen to satisfy a phase-margin requirement; the phase-margin specification is closely but not exactly related to the closed-loop frequency-response peak, which in turn is not a sure indicator of the transient response.

The energy expended in preparing a root-locus, gain-contour, or root-contour increases more rapidly than the number of poles and zeros of the open-loop transfer function but the effects of changing parameters are usually evident. The gain or other parameter can be set to yield a specified damping factor for a single pair of closed-loop complex poles or a specified value for a single real pole; the remaining poles are thereby fixed but their actual locations can be found only after additional analytical or graphical labor. Root-locus methods, by actually locating the closed-loop poles, most nearly specify the terms of the transient response. However, frequency response information, which is significant in its own right, is obtained only after additional computation.

The good designer, then, must have a working knowledge of all the available methods and benefits by extensions in the theory pertaining to any method. The current paper makes a valuable contribution to the root-locus method.

The general expression of the root-locus equation 8 does not require the introduction of h and the summation can be taken over all values of δ, even as well as odd, between 1 and h since the term $\sin(\delta\Pi)/2$ will automatically force undesired even terms out.

Akira Nomoto (Chuo University, Tokyo, Japan): The author expands the concept of the root locus and exhibits tables conveniently utilized for the practice of the method. He shows that the interchangeability between root locus and either root contour or gain contour are important as far as the problem is concerned with the analytical discussion. His treatment as to specific functions $T(3,0)$, $T(3,1)$, and $T(3,2)$ is to be generalized. For instance, if a transfer fuction $KY(s)(s+\alpha)$ of type $T(n,m)$ is the loop transfer function, the characteristic equation can be written as

$$1+\frac{\alpha KY(s)}{1+KsY(s)}=0$$

where $KY(s)/[1+KsY(s)]$ is of the $T(n,m-1)$ type. Gain contour $G(n,m;\alpha)$ corresponds to a root locus of transfer function $T(n,m-1)$. The details of this $T(n,m-1)$ function may be obtained from the former $T(n,m)$ function through the root-locus calculation for arbitrary K. In this generalized root-locus method, a rule that root loci of $T(n,m)$ and $T(m,n)$ are interchangeable seems profitably utilized.

Evaluation of Radio Influence Voltages in Fluorescent Lighting Installations

F. H. WRIGHT
ASSOCIATE MEMBER AIEE

S. A. ZIMMERMANN
MEMBER AIEE

ANY electric device whose operation produces a very rapid build-up or interruption of electric energy can cause a disturbance that may adversely affect radio reception. The communicator of a d-c machine represents a mechanical means of producing radio-frequency (r-f) power from electrical transients. A fluorescent lamp represents an all-electric voltage breakdown type of transient that produces modulated r-f energy. If no effort is made to confine this energy, near-by radio sets may experience a poor signal-to-noise ratio. This paper primarily discusses the radio influence voltages associated with commercial and domestic fluorescent lighting systems. The principal subjects considered are

1. The mechanism of r-f energy coupling between the lighting system and the radio.

2. A test method that makes possible the assignment of an approximate numerical value to the probable radio influence voltage of a fluorescent lighting system.

3. A statistical means of critically evaluating the probable radio influence voltage of a fluorescent lighting system.

Paper 56-75, recommended by the AIEE Production and Application of Light Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Winter General Meeting, New York, N. Y., January 30-February 3, 1956. Manuscript submitted April 1, 1955; made available for printing November 18, 1955.

F. H. Wright and S. A. Zimmermann are with the General Electric Company, Fort Wayne, Ind.

Coupling Between Lamp System and Radio

To cope with the unwanted r-f energy, it is necessary to consider the energy radiated directly from the lamps and fixture, the energy conducted to the radio directly through the power circuitry and the energy radiated from the wiring that supplied the power to the lighting system. The last-named energy is usually the most important of the three components. Direct radiation from the lamps and fixture can readily be disposed of by stating that the radio antenna should be physically remote[1] (10 feet is a conservative distance) from these objects, and that the fixture should be properly grounded. The remainder of

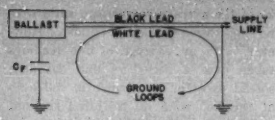


Fig. 1. Radiation loops of power input wiring

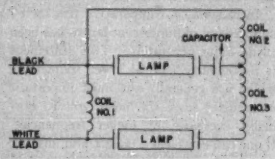


Fig. 2. Schematic of series ballast circuit diagram[4]

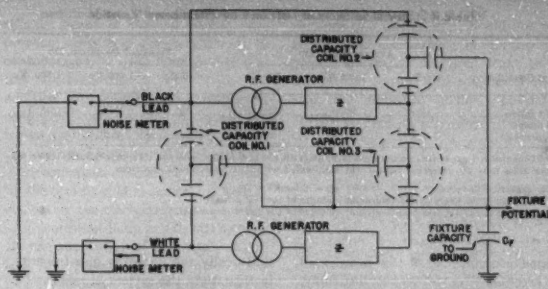


Fig. 3. Descriptive circuit diagram at radio frequency of a series ballast

this paper is primarily concerned with supply line radiation and radio influence voltages conducted directly to the radio. They are so interlinked that it is unnecessary to treat them separately. In some cases it may be necessary to consider the possibility that radiated energy from the lamps will couple with, and then re-radiate from the wiring.

Means of Evaluating Suppression

It is very important to realize that the power wiring to the ballast, normally thought of as a 2-wire system, is in reality a 3-wire system at radio frequencies as shown in Fig. 1. The third conductor is ground. In fact, if one neglects the ground circuit, he neglects, as will be shown later, practically the entire radio noise problem. This is true whether the lighting fixture and ballast case are grounded or ungrounded. The noise evaluation scheme presented herein neglects the line-to-line circuit, and considers only the line-to-ground circuits.

A radio noise meter[3] is useful for indicating the existence of radio influence voltage across various terminals in a fluorescent lighting system. These voltages are caused by the capacitive coupling between the various ballast coils and case. Fig. 2 shows the schematic circuit diagram for a series 2-lamp ballast. To illustrate how capacitive coupling to case might cause line-to-ground voltages Fig. 3 shows a crude lumped-constant r-f equivalent circuit for the ballast of Fig. 2. It is not intended to be an accurate description of the r-f circuit. Actually, the circuit varies with frequency and contains inductive as well as capacitive coupling between coils. For the purpose of description, the intercoil capacity and the coil-to-fixture capacity are assumed to result in a wye connection of three capacitors. However oversimplified this r-f equivalent circuit may be, it does serve to help visualize the effects of grounding various parts of the system. For example, an ungrounded fixture is associated with the ground loop circuit by its capacity to ground. Therefore, to ground the fixture (short-circuit C_f) may very significantly increase the effects of conducted noise because the capacity to ground is short-circuited which decreases the loop impedance and results in increased r-f currents in the loop.

Despite the fact that the line-to-ground voltages occur by devious means, they are of primary importance. The ground loop, because it has more area, represents a more effective radiating loop antenna than does the loop between lines. Furthermore, if both line-to-ground voltages are negligible, the line-to-line voltage is automatically negligible since the sum of the three must equal zero. For these reasons the test procedure about to be described considers only line-to-ground voltages.

An important stumbling block in a radio influence voltage evaluation of this kind is the variability of the r-f properties of the fluorescent lamps. Five to one changes in radio influence voltages in a short period of time may be expected.[3] In comparing different filtering schemes on a given ballast, these variations are not a serious problem as long as the suppression devices can be introduced and removed in such a way that the lamp behavior can be observed at all times. Several trials readily produce a fairly good measurement of the attenuation introduced by the suppression scheme, provided that the lamps have not changed; if they have changed, one can discount those data.

When the filtering is built into the ballast, the problem of determining the attenuation is more formidable. One solution is the introduction of a stable noise source to replace the lamps, but this is a scheme that can yield unrealistic results. If the lamp noise output varies in the laboratory, it will also vary in practice and should be considered. Furthermore, it is difficult to reproduce the characteristics of the lamps exactly. The capacitive and inductive coupling to ballast leads and the fixture should be the same and the variable generated energy and internal impedance of the lamps should be duplicated as they

Table I. Typical Data of New Lamps as a Function of Time (Tests Taken Over 1/2 Hour or Longer Intervals)

Test No.	600 Kc				1,000 Kc				1,500 Kc			
	Line A		Line B		Line A		Line B		Line A		Line B	
	Ref.	Test	Ref.	Test	Ref.	Test	Ref.	Test	Ref.	Test	Ref.	Test
1	76	785	31	885	24	240	13	240	17	78	6.5	82
2	117	717	53	1,140	43	417	22	453	29	103	10.5	108
3	238	3,370	95	5,240	87	1,220	48	1,940	60	235	24	385
4	205	4,000	85	4,350	57	655	30	750	44	170	15.5	162
5	100	1,540	65	1,540	53	615	33	665	126	315	53	305
6	700	5,740	285	5,500	317	2,100	165	2,100	165	505	73	475
7	445	5,640	178	5,500	220	1,820	110	1,800	133	445	59	435

All data are in microvolts and are given as the average of the two line-to-ground readings. The same 96-inch T12 lamps were used throughout these tests.

Table II. Statistical Summary of Data on Fluorescent Lamp Variation

Test Condition	Average Ratio, Db			Standard Deviation, Db		
	600 Kc	1,000 Kc	1,500 Kc	600 Kc	1,000 Kc	1,500 Kc
New lamps	23.6	22.3	11.5	3.1	3.7	5.0
All lamps	23.4	23.7	19.0	3.3	5.5	9.9
Suppressed*	−4.2	4.1	2.4	2.2	2.2	4.4

* 0.01-microfarad capacitors across each lamp in both the test ballast and the reference ballast. For all other data, only the reference ballast was suppressed with 0.01-microfarad capacitors.

Data expressed in voltage db = 20 log$_{10}$ V_T/V_R
V_T = average is the average line-to-ground voltages of the test ballast
V_R = average is the average line-to-ground voltages of the reference ballast

normally occur. All of this is awkward to accomplish or to neglect justifiably, especially since neither is necessary. A reasonably successful solution to this problem is to compare the ballast to be evaluated with a reference ballast or condition. This is done by rapidly switching the lamps from the reference to the test ballast and comparing the line-to-ground voltages. Fig. 4 illustrates the switching circuitry. As a comparison figure, the ratio of the average line-to-ground voltages of the two ballasts expressed in decibels (db) is used. Expressed mathematically, the comparison ratio in db = 20 log$_{10}$ V_T/V_R where V_T is the average line-to-ground voltages of the test ballast and V_R is the average line-to-ground voltage of the reference ballast. The average includes readings from both lines to ground and several sets of readings should be considered for each frequency evaluated. The standard National Electrical Manufacturers Association 600-ohm network[3] or some other reference may be used to simulate the r-f impedance of the power lines.

Statistical Evaluation of Probable Field Performance

The variation of the r-f characteristics of fluorescent lamps with time and between lamps appears to be nearly random in nature. That it is not purely random was indicated first by the fact that interference complaints could sometimes be solved by replacing old lamps. A study of several lamps over a long period of time was made to evaluate statistically the variation of r-f energy. In this study the system of comparison described herein was used. The reference ballast had 0.01-microfarad capacitors across each lamp it supplied. The test ballast contained no suppression except for the data summarized as "suppressed" under test condition. All the data were observed about 10 seconds after the lamps were switched from one ballast to the other, to avoid transient effects due to the switching. Readings were taken at three frequencies, 600 1,000, and 1,500 kc, to cover the spectrum of commercial amplitude-modulated broadcasting frequencies. The data were grouped into two time intervals: a short interval and a long time interval. The data of Table I are a typical sample of long time intervals. Each of these pieces of data represents the average of three short-interval pieces of data. The long interval had a duration of 1/2 hour to 2 hours. The short interval was about 1 minute. The long time interval was not stated specifically since the data were to be taken at random.

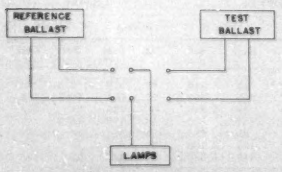


Fig. 4. Switching circuitry

Table II summarizes the information obtained from an investigation of the variability to be expected in the comparison ratio due to lamp variation. These data represent information on seven sets of 96-inch *T12* fluorescent lamps. Of the 123 pieces of data taken at each frequency, 21 were associated with very noisy lamps near the end of lamp life. These lamps actually failed during the test and, as shown by the standard deviation of all lamps compared to new lamps, they increased the observed variability. It is evidently advisable to use new lamps in this comparative test. There is also a definite indication that the suppressed (0.01-microfarad) lighting system exhibits less variation than the unsuppressed system.

The data of Table II indicate that (with a confidence limit of two standard deviations) 95 per cent of all observations taken over a short period of time will fall within ±6.2 db at 600 kc, ±7.4 db at 1,000 kc, and ±10 db at 1,500 kc of the average comparison ratio with no suppression on the test ballast. However, with an average of five readings taken over a period of one day, the spread expected by chance is ±2.8 db at 600 kc, ±3.3 db at 1,000 kc, and ±4.5 decibels at 1,500 kc with the same degree of confidence. The lamps are left on for the entire day of testing.

In this way a comparative test for r-f suppression of fluorescent lamps has been made quite accurately. This statement is particularly true at lower frequencies which are generally the most important frequencies since the noise voltage normally decreases at higher frequencies. If one can establish a ballast and filtering scheme that has given satisfactory field performance, he can state from the foregoing statistical analysis that any new ballast must have average conducted noise voltages less than 5 db above the selected standard at each of the frequencies mentioned. In fact, it would be fairly safe to require only a 600-kc reading and a maximum ratio of +3 db for a satisfactory limit.

Fig. 5 shows a plot of additional data taken on 48-inch fluorescent lamps. The graph on the left compares two similar ballasts, one (the reference) with 0.01-microfarad capacitors across the lamps and the other with no suppression. In the right-hand chart, the test ballast has a 0.01-microfarad capacitor across the line. The spread in the short-time readings is indicated by the length of the vertical lines. The average of five of these readings is shown by a short horizontal line. The average data for the test ballast suppressed by a single 0.01-microfarad capacitor across the line fall within the 5-db limit established by the previous investigation. Therefore,

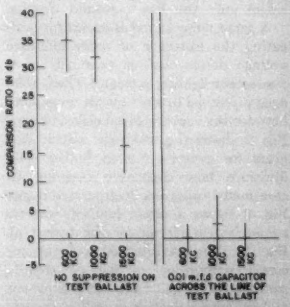


Fig. 5. Comparison evaluation of a 48-inch fluorescent lamp system (the reference has a 0.01-microfarad capacitor across each lamp)

the new suppression scheme is in this case adequate. It is also interesting to note that the spread in the data shown in Fig. 4 for short-interval readings falls within the spread for these readings established by the 95-per-cent confidence limit of the data of Table II.

Conclusions

The apparent way to filter high-frequency power from a system is to shunt the power through a lower impedance to ground. As a general procedure on fluorescent lamp installations, it is a questionable procedure. Ground can frequently become a nebulous quantity. Private homes may be particularly troublesome in this respect. All the data the authors have observed suggest capacitors across the lamps as the most reliable means of internal ballast suppression. For this reason the use of capacitors across each lamp in the system has been used as a reference. In addition to reducing and stabilizing conducted noise, capacitors across the lamps may reduce direct radiation from the lamps and fixture by as much as 25 db.

A comparison scheme may successfully fill the present void in general methods of evaluating the conducted noise potential of a fluorescent lamp-and-ballast system. By the application of elementary statistics, it is possible to make quite effective evaluations. The suggested filtering scheme for the reference is the application of 0.01-microfarad capacitors across each lamp in the system.

In specifying a 5-db limit on the difference between a filtered ballast and a reference ballast, it was assumed that the variance of a filtered system will not be greater than the variance of the unsuppressed system. It has been shown that the filtered variance is sometimes less than the unfiltered ballast. To date greater variance has not been detected.

References

1. Suppression of Radio Interference in Fluorescent Lighting Installations, Glenn Walters. *Illuminating Engineering*, New York, N. Y., vol. XLIX, no. 6, June, 1954, pp. 295–300.

2. Methods of Measuring Radio Noise. Joint Edison Electric Institute, National Electrical Manufacturers Association and Radio Manufacturers Association Report. *NEMA Publication no. 107*, National Electrical Manufacturers Association, New York, N. Y., 1940, 10 pp.

3. A Preliminary Study of Radio Interference as Caused by Fluorescent Lamps in the Home, L. F. Shorey, S. M. Gray. *Illuminating Engineering*, Mar. 1947, pp. 365–76.

4. Apparatus for Starting and Operating Gaseous Discharge Devices, Albert E. Feinberg. *U. S. Patent 2,558,293.*

Discussion

C. W. Frick (General Electric Company, Schenectady, N. Y.): One of the difficulties encountered in determining the influence of fluorescent lamps on radio reception has been the spread between measurements on different lamps even though the lamps are apparently identical. Such variability may be expected in devices which include spark gaps or gaseous conductors. It is well known that a commutator motor exhibits greater variability of radio noise when sparking is present than when it is absent. The authors suggest a useful test procedure whereby the effect of such variability is minimized by making comparisons with a reference circuit using the same lamp.

The paper makes an interesting observation that the variability increases with the frequency at which the measurements are made, over the range of 600 to 1,500 kc. I do not believe such an observation has been reported before. Examination of the results of some tests I made several years ago, together with some similar tests from other sources shows the same trend. In all these tests radio noise measurements on three lamps of the same type were compared. I computed the spread between lamps as a percentage of the average and found that except in some few cases where one of the readings appeared to be erroneous, this percentage spread consistently increased from 600 to 1,500 kc. However, the increase was not very marked between 600 and 1,000 kc, which seems to agree with the observations on new lamps in Table II. This suggests that readings be taken at 1,000 kc or under unless there is reason for doing otherwise.

There may be some connection between the increase of variability with frequency and the duration of short pulses or time in which sudden changes take place in the lamp. To show this we can make use of the familiar mathematical concept that a variable voltage or current may be represented either by a function of time or by a function of frequency which is called a spectrum. Our measuring instrument, the radio noise meter, takes a narrow-band sample of the spectrum, the position on the spectrum being determined by the frequency to which the instrument is tuned. Other things being the same, the instrument reading is proportional to the ordinate of the spectrum curve.

Our experience indicates that the spectrum curve of a fluorescent lamp (instrument-reading plotted against frequency) is fairly flat between 500 kc and 1 or 2 megacycles and then falls off rapidly. It can be shown theoretically that the spectrum for a pulse 0.3 microsecond wide is approximately flat up to 1 megacycle and then falls to zero at about 3 megacycles. Pulses would be expected of about that width produced in fluorescent lamps or wave fronts of about that steepness which would have somewhat the same effect on the shape of the spectrum. This time element may be variable between, say, 0.1 and 1.0 microsecond in which case the spectrum above 1 megacycle would be subject to considerable variation even though it is quite steady below that point. This would seem to be a reasonable explanation of some of the effects observed in the various tests.

The paper suggests an equivalent circuit for the lamp and associated ballast; see Fig. 3. The sample taken from the spectrum is treated as if it were a single frequency, an assumption which has been found reasonable in practice. The lamp is represented as a r-f generator acting through an impedance Z. In our tests this impedance was found to be equivalent to a resistance. Changes in the spectrum such as those mentioned change either the equivalent source voltage or the equivalent impedance or both. When the voltage and the impedance do not change, the effect of changes in the associated circuit can be determined. The circuit as shown in Fig. 3 is not quite complete in that the supply circuit or house wiring is not included. The supply circuit should be represented by an impedance from the black lead to ground and one from the white lead to ground. The noise meter should be represented as a voltmeter across either of these impedances. Then the effect of short-circuiting capacitance C_F is apparent if it is assumed that the r-f generator does not change.

F. H. Wright and **S. A. Zimmermann**. The authors are indebted to Mr. Frick for his explanation of lamp variability with frequency. It supplies a better understanding of the data. Such an analysis was beyond the scope of this work.

The effective r-f impedance of the line should have been included in Fig. 3. The 600-ohm National Electrical Manufacturers Association network was used in the reported experiments.

The authors did notice that variations in associated circuitry could change the lamp r-f characteristics, e.g., placing capacitors across the lamp. However, the process of short-circuiting the fixture capacity to ground did not appreciably affect the lamp characteristics. The behavior was the same as that predicted by the described equivalent circuit.

Test and Operating Experience with the Triple-Diametric Rectifier

R. V. WACHTER
MEMBER AIEE

C. S. HAGUE
ASSOCIATE MEMBER AIEE

C. R. MARCUM
MEMBER AIEE

THERE is an ever-present demand for improvements in electric equipment. To improve the equipment, the engineer may try many new ideas. Some of these will yield results while others will be discarded as not being worth while. On equipment such as mercury-arc rectifiers many ideas have been successful and the improvements have been frequent and substantial. Usually, the engineer has so many new ideas to try that he does not take time to re-examine old ideas in the light of recent developments.

Approximately 30 years ago, the triple-diametric circuit was used with multianode mercury-arc rectifiers in Europe. For the multianode rectifiers and for the type and size of installations made at that time, other power circuits were more advantageous. The use of the triple diametric circuit was discarded.

During the last 20 years, the design and application of mercury-arc rectifiers in this country have undergone radical changes. The single-anode designs have completely outmoded the multianode designs. The aluminum industry with its rapid growth has required larger and larger blocks of d-c power.

Over the last decade, the rating of a single rectifier unit has been increased by 60 per cent. At the same time, the size of single loads in the aluminum industry has increased 150 per cent. With single loads of 100,000 kw supplied by parallel rectifier units, there can be no question of the need for increased ratings of rectifier units. This paper describes a practical re-examination of a discarded idea, the triple-diametric circuit, in the light of interim developments.

Comparison of the Triple-Diametric and Double-Wye Power Circuits

The 6-phase double-wye rectifier power circuit is used on practically every large rectifier installation in this country. If the triple-diametric circuit offers any over-all improvement, the improvement is relative to the characteristics of the double-wye circuit.

Fig. 1 shows the practical parameters of the triple-diametric and the double-wye circuits when applied to a 6,000-ampere 750-volt rectifier unit. The need for increased unit ratings is to be considered in examining these parameters. A summary examination of Fig. 1 will show the favorable and unfavorable parameters of the triple-diametric circuit to be as follows:

Favorable: lower tube peak current; lower rms secondary phase current; and lower arc-back current.

Unfavorable: higher tube inverse voltage; higher d-c winding voltage; and larger transformer size (including interphase).

To evaluate this summary examination, it must be viewed through the eyes of the equipment designer. The limiting factors to increasing the rating of single-anode tubes are usually the thermal capacity and the arc-back rate. The heating in a tube is approximately proportional to the average current through it. For a given number of tubes and a given rating, the average current per tube is the same for all power circuits. It is therefore impossible to increase the rating of a tube beyond its thermal capacity by changing power circuits. In general, the tubes applied in units at 600 volts d-c and above are limited in rating by the arc-back rate. For a given commutation angle, the probability of arc back is a function of peak current in the tube and the inverse voltage applied to the tube, if other conditions are unchanged. Since tubes are designed to apply over a range of voltages, it is very possible that, for some voltage ratings, the probability of arc back is much more a function of the peak current than of the inverse voltage. For these cases, the triple-diametric circuit offers a means of obtaining increased ratings from single-anode tubes.

Undoubtedly, the major reason that this circuit was not applied to multianode rectifiers was that the total direct current must still flow in the same vacuum envelope. Therefore, the peak current in the envelope is the same regardless of the power circuit.

To increase the rating of a transformer by design is usually considered very simple. However, two problems to be considered are the manufacturing difficulties of winding large cross-sectional area conductors and the bracing of conductors to withstand the forces caused by larger fault currents. The triple-diametric circuit results in a lower rms secondary phase current and a lower arc-back current which helps to overcome both of the problems. The over-all size of a transformer for a triple-diametric circuit is somewhat larger, primarily because of the interphase, which is an economic consideration against the use of this circuit. The increased secondary voltage of the triple-diametric circuit is not a problem in the transformer design since other considerations determine the voltage insulation level.

The d-c circuit-breaker equipment ratings are not affected. Some of the rating factors for anode circuit breakers are the rms current through the breaker in normal operation, the crest fault current in the breaker, and the voltage applied across the contacts immediately after a fault is cleared. The reduced rms secondary current and the reduced arc-back current of the triple-diametric circuit are favorable for an increased rating for the anode circuit breaker. The increased secondary voltage is unfavorable. As in the case of the tubes, breakers are designed to apply over a range of voltages. For many applications, the increased secondary voltage will be below the rating of the breaker.

Having compared the parameters of the triple-diametric and double-wye power circuits from a designer's viewpoint, it can be concluded that, for some applications and ratings, the triple-diametric circuit offers a very good possibility of obtaining increased unit ratings.

Laboratory Tests

As a first check on the possibility of obtaining increased ratings, tubes were load-tested in the laboratory at direct voltages for which it was known that the rating was limited by arc-back rate. With the same tubes tested on both types of circuits, the data showed that an increase in rating of approximately 20 per cent could be expected.

Paper 56-186, recommended by the AIEE Industrial Power Rectifiers Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Winter General Meeting, New York, N. Y., January 30-February 3, 1956. Manuscript submitted October 19, 1955; made available for printing November 23, 1955.

R. V. WACHTER is with the Aluminum Company of America, Wenatchee, Wash.; C. S. HAGUE and C. R. MARCUM are with the Westinghouse Electric Corporation, East Pittsburgh, Pa.

CIRCUIT	DELTA, SIX PHASE TRIPLE DIAMETRIC	DELTA, SIX PHASE DOUBLE WYE
RATED D-C VOLTAGE E_d	750 VOLTS	750 VOLTS
RATED LOAD CURRENT I_d	6000 AMPS.	6000 AMPS.
LOAD POWER P_d	4500 KW	4500 KW
PEAK ANODE CURRENT	2000 AMPS.	3000 AMPS.
AVERAGE ANODE CURRENT	1000 AMPS.	1000 AMPS.
RMS ANODE CURRENT	1414 AMPS.	1732 AMPS.
D-C WINDING VOLTAGE E_s	883 VOLTS	680 VOLTS
INITIAL INVERSE VOLTAGE E_{ii}	1153 VOLTS	772 VOLTS
PEAK INVERSE VOLTAGE E_{pi}	2500 VOLTS	1670 VOLTS
COMMUTATING ANGLE U (NO DELAY)	27.5°	27.5°
CREST ARC-BACK CURRENT-ONE GROUP	35,700 AMPS.	61,200 AMPS.
TRANSFORMER WINDINGS RATING	6400 KVA	6040 KVA
INTERPHASE EQUIVALENT RATING	870 KVA	320 KVA

Fig. 1. Comparative chart of the parameters of the triple-diametric and the double-wye power circuits

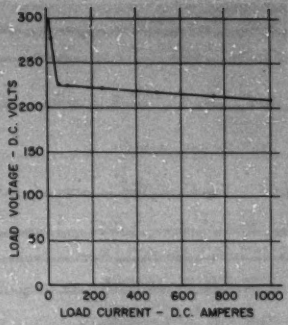


Fig. 3. Voltage-regulation curve for a rectifier unit rated 1,000 amperes, 200 volts d-c

A 1,000-ampere 200-volt rectifier unit was installed in the laboratory for a further study of the characteristics of the triple-diametric circuit. Fig. 2 shows the anode current of three tubes when the unit was operating at full load with no phase control. Fig. 3 shows a voltage-regulation curve for this unit. From this curve, it can be seen that, when the unit reverts to 6-phase operation at very light loads, the direct voltage rises 50 per cent. On electrochemical applications, the units are always heavily loaded so the light-load voltage rise presents no problem. On other applications, the voltage rise can be prevented, as it is on the double-wye circuit, by a small dummy load resistor. Concurrent with the model tests in the laboratory, equipment was manufactured and installed for field tests.

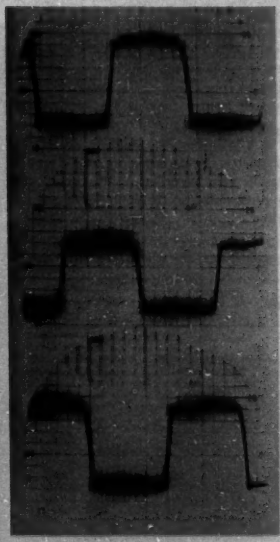


Fig. 2. Anode currents for a rectifier utilizing the triple-diametric power circuit

Field Tests and Operating Experience

The reactive voltage drop of a rectifier unit of a given output rating is expressed by the equation

$$E_x = \frac{pI_{do}X_c}{2\pi n}$$

where

p = the number of phases in a commutating group
n = the number of commutating groups per unit
X_c = the commutating reactance per phase, ohms
I_{do} = d-c output current

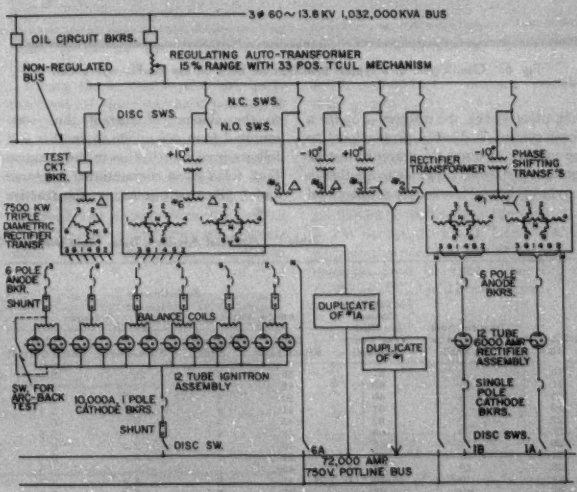


Fig. 4. Single-line diagram of 72,000-ampere installation showing experimental circuits

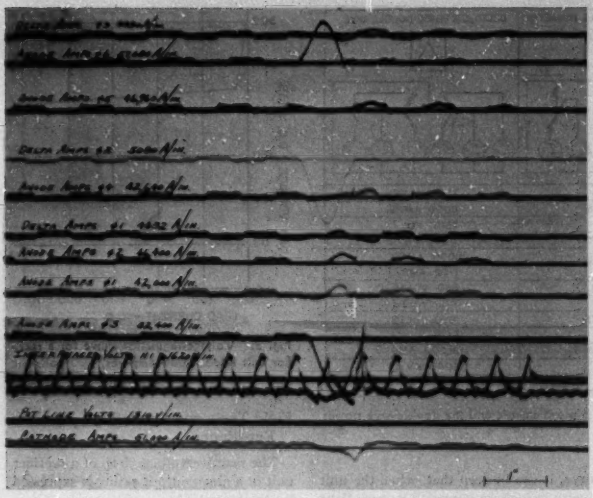


Fig. 5. Oscillogram of a typical transition arc back on the triple-diametric circuit

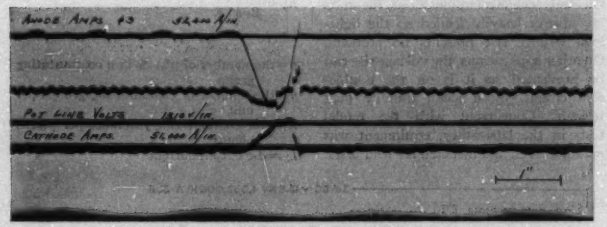


Fig. 6. Oscillogram of a typical transition arc back on the double-wye circuit

In other words, the reactive drop for a given current is directly proportional to the number of times the current must be commutated from one anode to another in a given cycle, and is inversely proportional to the amount of current thus commutated from one anode to another. It follows from substitution in the equation that, for a given commutating reactance and a given output rating, the reactive drop of a unit utilizing a triple-diametric circuit is only 4/9 that of a unit utilizing the conventional double-wye circuit. Expressed from another viewpoint, for a given output voltage regulation and a given rating, the commutating reactance for a triple-diametric unit can be approximately twice that for a double-wye unit. This reactance is the major limiting factor of fault current when an arc back occurs. Consequently, lower arc-back currents are to be expected from any practical installation utilizing a triple-diametric circuit.

To verify this comparison of arc-back current magnitudes, a 10,000-ampere 750-volt transformer, connected delta 6-phase triple-diametric, was placed in parallel with six 6-phase parallel double-wye units comprising a 72,000-ampere installation at the Wenatchee Plant of the Aluminum Company of America. A 12-tube rectifier assembly was arranged so that it could be connected through a 6-pole anode circuit breaker to either the secondary terminals of the triple-diametric transformer or to the terminals of one double-wye secondary group of an existing transformer. Fig. 4 is a single-line diagram showing the electric circuits for the installation and the manner in which the equipment to be tested was connected.

To perform the tests, a short-circuiting switch was connected in parallel with the rectifier tubes on phase 3 of the unit. Data were recorded by means of a magnetic oscillograph. Arc backs were initiated by closing the short-circuiting switch from a cam on the oscillograph drive. Fig. 5 shows a typical oscillogram for a transition arc back on the triple-diametric circuit, while Fig. 6 is a typical oscillogram recording fault circuit and potline data for the double-wye connection.

Numerical values of the data from the oscillograms are shown for the triple-

Table I. Artificial Arc-Back Data, Triple-Diametric Circuit

Oscillogram No.	Degrees from Transition to Arc Back	Potline Volts During Arc Back (Per Cent of Initial)	Fault Current: Anode, Phase 3		Fault Current: Cathode		Crest Anode Current, Kiloamperes: Phases					Crest A-C Winding Current as a Percentage of Rms Rating: Phases		
			Crest, Kiloamperes	Maximum di/dt	Crest, Kiloamperes	At Crest Anode Current	1	2	4	5	6	1	3	2
16	0	67	50	8	21	6.3	7.5	7.6	6.5	9.4	39	270	2700	231
17	0	66	48	8	15	0	7.1	7.9	6.0	6.0	40	300	2940	248
18	0	69	50	8	17	4.2	8.4	8.0	5.9	6.6	37	198	2670	246
19	0	67	51	8	13	5.1	8.4	8.1	6.4	9.4	37	216	2580	270
22	0	67	45	8	10	0	6.3	8.0	7.5	9.3	41	216	2700	300
27	0	66	48	8	11	2.5	6.7	6.4	5.3	8.9	40	216	2700	237
20	170	67	40	8	2.5	0	8.4	8.1	5.3	5.9	35	222	2350	210
21	70	66	24	7	4	0	4.6	4.8	4.5	4.7	24	171	1470	192
23	60	74	42	7	6	6.3	8.4	7.5	4.5	6.0	28	222	2070	300
24	40	70	41	8	6.3	6.7	8.5	7.0	4.3	5.9	27	222	2070	300
25	40	65	27	7.5	5.0	5.0	6.3	5.8	5.5	5.3	16	222	1230	300
26	50	65	33	7.5	5.1	5.0	9.4	6.5	5.3	4.8	24	222	1380	300

Table II. Artificial Arc-Back Data, Double-Wye Circuit

Oscillogram No.	Degrees from Transition to Arc Back	Potline Volts During Arc Back (Per Cent of Initial)	Fault Current: Anode, Phase 3 Crest, Kiloamperes	Fault Current: Anode, Phase 3 Maximum di/dt	Fault Current: Cathode Crest, Kiloamperes	Fault Current: Cathode Maximum di/dt	Crest Anode Current, Kiloamperes: Phases 1	2	4	5	6	Crest A-C Line Current as a Percentage of Rms Rating: Phases 1	2	3
33	0	60	59	8.8	41	7.5	20	5.0	..	20	..	993	967	380
34	0	59	60	8.8	41	7.5	21	5.0	..	21	..	993	967	453
39	75	60	44	11	34	9.5	19	6.5	..	9.4	..	814	636	229
35	110	58	34	10	24	6.0	18	5.5	..	0	..	686	495	178
36	180	67	48	8.8	33	5.0	17	5.0	..	18	..	762	788	356
37	30	58	52	8.8	39	5.0	20	5.5	..	17	..	916	787	305

diametric circuit in Table I and for the double-wye circuit in Table II. In comparing numerical values of Tables I and II, it should be remembered that the triple-diametric unit is rated 10,000 amperes and the double-wye unit is rated 6,000 amperes. Table I lists, in addition to the maximum value of reversed cathode current, its value at the instant the fault current in the anode circuit reaches its first crest, as there is significant difference in these two values of cathode current for triple-diametric operation. On both tables, data on alternating current and crest current in the other secondary phases are listed.

In normal operation, most arc-backs occur at the end of the conduction period at the point of transition between the conducting and nonconducting intervals. Furthermore, both theory and experience show that the fault-current magnitude from arc backs which do occur at this time is greater than that from arc backs which occur at other periods of the cycle. Therefore, in evaluating the data, primary consideration is given to those pertaining to transition faults.

Of the several tests made on the double-wye circuit, only two represented transition conditions. It was felt that these two tests were sufficient to draw comparisons for two reasons. First, the data for the two tests were duplicate within practical accuracies of reading the oscillograms, and similar previous tests have indicated a repetition of duplicate fault-current magnitudes for transition arc backs. Second, the value of fault currents of these two tests was practically the same as that obtained on previous tests of an installation of essentially duplicate size and characteristics.

In comparing the data for the two circuits, it is immediately evident that with the triple-diametric unit very little d-c fault current is present in the faulted anode circuit at the time the first crest value is reached. A visual comparison of the cathode current trace for the double-wye circuit, shown in Fig. 6, with that for the triple-diametric, shown in Fig. 5, vividly illustrates this point. With the d-c contribution thus minimized, the total fault current at the time when high-speed anode circuit breakers become effective is appreciably reduced. Taking into account the difference in the rating of the two transformers tested, the maximum secondary fault current for the triple-diametric unit is approximately 21 times normal rms rating, whereas this figure is 35 for the double-wye circuit.

During the first cycle after arc back, reactance of the interphase winding is effective in minimizing the flow of current from other parallel groups to the faulted phase for both the double-wye and triple-diametric modes of operation. Thus, with double-wye operation, two anodes displaced 120 degrees from the faulted phase experience high forward currents while, with triple-diametric operation only one anode displaced 180 degrees from the faulted phase has high forward currents. Considering the difference in rating of the two units, the data show the magnitude of this forward current to be 20 per cent higher for the triple-diametric circuit. This higher current carried by the diametrically connected anode will cause an increase in the incidence of sympathetic arc back.

In comparing the magnitude of current in the a-c circuits, it is important to remember that Table I gives the percentage of normal primary winding current, whereas Table II lists the percentage of normal line current. Since the line current for the double-wye circuit is divided between two sets of primary windings, the current in the primary associated with the faulted secondary reaches relative values approximately double those listed in Table II. Thus, the fault currents in the primary winding for the triple-diametric circuit are approximately 50 per cent greater than those for the double-wye circuit. This higher magnitude of primary current is within normal bracing design for rectifier transformers.

The evidence of these tests supports the claim that arc-back current is appreciably reduced through the use of the triple-diametric circuit. This will permit the use of fewer units of larger capacity to accomplish a particular potline rating, or it will permit the use of larger potlines without the need of external reactors to maintain fault currents within the capacity of available d-c switchgear.

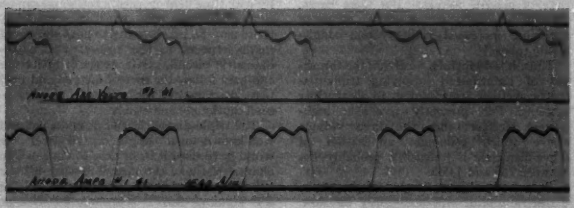


Fig. 7. Oscillogram of anode current and arc voltage with triple-diametric mode of operation

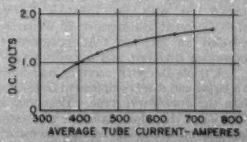


Fig. 8. Curve of difference in average arc voltage for the triple-diametric and double-wye modes of operation

Since the instantaneous anode current in the triple-diametric circuit is only 2/3 of that in the double-wye circuit for a given value of cathode current, the rectifier average arc drop should be reduced. Lower arc drop, of course, gives an operating advantage by increasing the over-all efficiency.

To verify this advantage, arc-drop measurements were made on the same rectifier tubes under actual potline operating conditions with both triple-diametric and double-wye modes of operation. These tests were made by the conventional oscillograph method and by the slightly modified version of the Swedish wattmeter method. Fig. 7 is a typical oscillogram with voltage and current traces. This oscillogram shows the typical phenomenon of decrease in arc voltage with duration of conduction. Thus, the arc drop at a given cathode current of triple-diametric operation should be further reduced because of the lower averaging effect of the additional 60 degrees of conduction.

Fifteen sets of data were taken. In each case arc drop was determined by wattmeter and oscillograph methods. The analysis of this data shows the arc drop of the triple-diametric circuit to be approximately 1.5 volts less than the double-wye for the same average anode current, as shown in Fig. 8. The arc drops are the same when the average load current for double-wye operation is 0.625 that for the triple-diametric operation.

These data indicate that over-all efficiency of a rectifier installation will be slightly improved through the use of the triple-diametric circuit, provided the tubes are applied at average currents of 150 per cent or less of those prevailing for double-wye operation.

At the completion of the arc-back and arc-drop tests in November of 1954, the 7,500-kw transformer was connected through a 10,000-ampere 6-pole anode breaker to one of the existing 12-tube rectifier assemblies which carries a nameplate rating of 6,000 amperes. The unit was then operated in parallel with existing equipment on the potline bus. No phase delay was utilized and the experimental unit shared load according to its inherent voltage output. This was such that load on the unit stayed between the limits of 9,600 and 10,200 amperes. In spite of the fact that the tubes were operated at approximately 167 per cent of their rating, satisfactory operation prevailed over this period with only a moderate increase in arc-back frequency from the unit.

Conclusions

Because fundamental equations indicate certain important practical advantages for the triple-diametric circuit particularly when used with single-anode rectifier tubes, laboratory and field experiments on equipment utilizing this circuit have been made to evaluate its performance. These experiments have shown the following important characteristics of the circuit as compared to conventional double-wye operation.

1. It is practical to build an effective 3-phase interphase transformer to enforce 180-degree tube conduction.
2. D-c contribution to arc-back current is greatly reduced for the initial period between the instant of arc back and the time high-speed anode circuit breakers effectively limit fault current.
3. Higher per-unit forward currents in the diametrically connected anode may cause an increase in the incidence of sympathetic arc back.
4. Tube arc voltage for a given average current is reduced.
5. Where tube ratings are normally limited by arc-back rate, average tube current can be appreciably increased without sacrifice in performance.

Although the circuit has some detrimental characteristics, these are not serious and its advantages for certain applications far outweigh the disadvantages. In general, the application of the circuit on large installations in electrochemical service will enable the accomplishment of a given d-c bus rating with fewer rectifier assemblies and associated equipment. This will result in smaller building space requirements, and reduced installation maintenance costs.

Because of these advantages and because of the reduction in magnitude of the arc-back current, which is a characteristic of this mode of operation, equipment utilizing this circuit is now being manufactured to furnish d-c power for three large electrochemical installations totalling over 200,000 amperes.

Discussion

A. Schmidt, Jr., L. W. Morton (General Electric Company, Schenectady, N. Y.): The various factory and field tests so well described in the paper cover the technical questions that could not be answered analytically on the basis of circuit theory. Primarily the objective was to determine how much the current ratings of certain well-developed rectifier tubes could be increased by use of the triple-diametric, as compared to the double-wye circuit, at voltage levels where ratings are limited by arc-back frequency. Test results, particularly those carried out with the 7,500-kw 750-volt triple-diametric transformer at Wenatchee, revealed an increase that rewarded the sponsoring engineers and organizations.

The triple-diametric rectifier transformer is a larger and more complicated transformer than the double-wye type. Economic application of the triple-diametric transformer depends on obtaining enough increase in output current for any specified direct voltage rating to more than offset the increase in transformer size and complexity. As far as the presently available well-developed and well-proved rectifier tubes are concerned, economic applications of the triple-diametric circuit are limited to lower direct voltages than the double-wye. A 50% higher peak inverse voltage for the triple-diametric circuit is the main cause of this limitation.

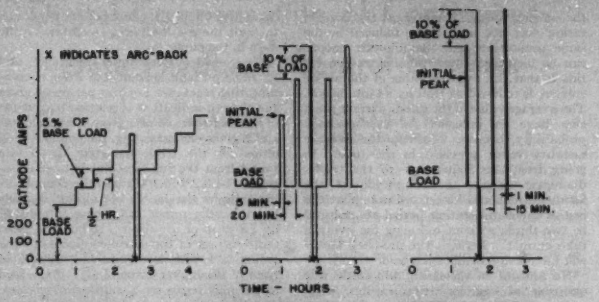
These facts were well recognized by the authors, for they state that, "Since tubes are designed to apply over a range of voltages, it is very possible that for some voltage ratings, the probability of arc back is much more a function of the peak current than of inverse voltage. For these cases, the triple-diametric circuit offers a means of obtaining increased ratings from single-anode tubes."

Thus, it is doubtful if we are on the threshold of immediately applying triple-diametric rectifier transformers to all rectifier systems. For presently available high-current rectifier tubes, the direct voltage above which economic application of triple-diametric is unfeasible probably lies between 750 and 1,000.

Progress in the application of the triple-diametric circuit will necessarily be slow. Exploratory tests should be made to determine accurately the upper limits of direct voltage where triple-diametric circuits can be safely and economically used for presently available high-current rectifier tubes. Investigation may prove that rectifier tubes of new design should be developed which can be used to economic advantage with triple-diametric rectifier transformers for the higher direct voltages where present tubes are unadaptable.

When mercury-arc rectifier ratings are limited by arc-back frequency, as is so often the case, experience has taught us that factory rating tests must be confirmed by long-time field operation before we can be sure of them. This paper indicates that there may be real benefits to both users and manufacturers in the use of the triple-diametric circuit. Safe progress in its use can only follow the sort of factory and field tests described in the paper. Let us proceed cautiously and wisely.

The triple-diametric circuit is an old and well-known one. Engineers all over the world have used it from time to time. One characteristic of the circuit is that all the tubes connected to any single phase of the d-c windings can be disconnected without appreciably reducing the rectified output voltage. This advantage led to its use in a



A. 30-MINUTE LOAD LIMIT
1. START AT BASE LOAD SPECIFIED
2. INCREMENT - APPROX. 5% OF BASE LOAD EVERY HALF-HOUR.
3. INCREASE LOAD UNTIL ARC-BACK OCCURS.
4. REDUCE LOAD 10% OF BASE LOAD AND CONTINUE AS BEFORE.
5. CONTINUE UNTIL 3 SUCCESSIVE ARC-BACKS ARE OBTAINED AT SUBSTANTIALLY EQUAL LOADS, INDICATING THAT LIMIT HAS BEEN REACHED.

B. 5 MINUTE LOAD LIMIT.
1. START AT BASE LOAD WITH VALUE EQUAL TO TWO THIRDS OF 30 MINUTE LOAD LIMIT.
2. INITIAL PEAK - 10% OF BASE LOAD GREATER THAN 30 MINUTE LOAD LIMIT.
3. DUTY CYCLE - 20 MINUTE BASE LOAD, 5 MINUTE PEAK, REPEAT.
4. INCREMENT - 10% OF BASE CURRENT ON SUCCESSIVE PEAKS.
5. IF ARC-BACK OCCURS, REDUCE LOAD 10% OF BASE LOAD AND CONTINUE AS BEFORE UNTIL THREE ARC-BACKS ARE OBTAINED AT LIMIT.

C. 1 MINUTE LOAD LIMIT.
1. START AT BASE LOAD WITH VALUE EQUAL TO TWO THIRDS OF 30 MINUTE LOAD LIMIT.
2. INITIAL PEAK - 10% OF BASE LOAD GREATER THAN 5 MINUTE LOAD LIMIT.
3. DUTY CYCLE - 15 MINUTE BASE LOAD, ONE MINUTE PEAK, REPEAT.
4. INCREMENT - 10% OF BASE CURRENT ON SUCCESSIVE PEAKS.
5. IF ARC-BACK OCCURS, REDUCE LOAD 10% OF BASE LOAD AND CONTINUE AS BEFORE UNTIL THREE ARC-BACKS ARE OBTAINED AT LIMIT.

Fig. 9. Procedure for load limit tests

line of sealed-tube rectifier units for essential elevator service in the early 1930's. Two-pole anode breakers, one for each of the three single-phase d-c windings, were used. In case of complete tube failure due to arc back, the anode breaker associated with the defective tube remained open after several reclosures. The remaining tubes continued to supply uninterrupted service to the elevators at nearly normal voltage.

To return now to technical details of the paper, these are a few comments and questions which may be helpful:

Examination of Tables I and II confirms analytical predictions that the fault current in the a-c windings of the triple-diametric transformer during arc back are higher on a per-unit basis than for the double-wye. This is of concern to transformer designers and must be factored into the over-all picture. Increased bracing will be necessary, which may or may not be prohibitive, depending on cases. The higher a-c line currents during arc back should not be overlooked.

Depending on the application and the degree of increased anode breaker maintenance, the increased tendency towards sympathetic arc back in the diametrically connected anodes, mentioned in the authors' conclusions, may be of concern. It is good that the authors noticed this possibility, for we must keep it in the back of our minds when applying the triple-diametric circuit.

The authors indicate that with equal values of regulation in the two circuits, the triple-diametric transformer has more than double the ohmic value of reactance in the double-wye transformer. A clearer picture of potential severity of arc-back current is given by the per-unit transformer reactance. For equal regulation, the triple-diametric transformer has only 70% of the reactance in the double-wye transformer on a per-unit base. Thus, the stresses in the triple-diametric transformer due to arc back without d-c feed would be double the stresses in the double-wye transformer. The saving factor in the triple-diametric circuit is the delay in build-up of d-c feed to the arc back. This delay represents the time required to saturate the interphase transformer. The oscillogram in Fig. 5 indicates the delay to be in the order of 135 degrees. This is more than double the value to be expected with a normal zig-zag interphase transformer design. We would appreciate the authors' comments on this point.

The project and investigation were ingenious, well planned, and courageously carried out.

C. C. Herskind (General Electric Company, Schenectady, N. Y.): The paper sets forth the more important considerations involved in the circuit analysis. The triple-diametric circuit possesses both advantages and disadvantages when compared with the double-wye circuit. There is no question about its being a practical circuit. The question is do the advantages outweigh the disadvantages, and what are the economic gains?

It is too early to make a final assessment of the gain which may be realized through the use of the triple-diametric circuit. These gains depend to a large degree on the amount by which the rating of a given complement of rectifier tanks may be increased over the rating obtainable with the double-wye connection. It may be worth while to re-examine the procedure in rating rectifier tanks in order to see how the difference between the two circuits must be evaluated.

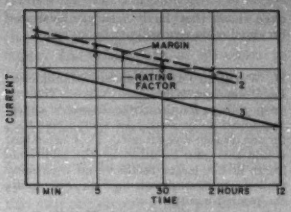


Fig. 10. Procedure for rating

Curve 1—Failure on load limit tests; (X) one or more arc backs in three trials
Curve 2—Capacity on load limit tests; (O) carried three times without arc back
Curve 3—Rating: performance in service, arc-back rate for rectifier unit, 2 to 3 ± 10 × per year

It is impractical to make adequate load tests at the factory to determine the rating of a rectifier. One procedure which is used for determining ratings is as follows:

1. Load limit tests are made at the factory using the procedure outlined in Fig. 9. These are accelerated tests in which the rectifier is loaded until it fails by arc back.

2. The rectifier capacity established by the load limit tests is then plotted on a chart as shown in Fig. 10. Curve 1 shows the currents at which arc-back failures occur. Curve 2 shows the loads that were successfully carried, i.e., without arc back. It is of interest that on many tanks this curve is practically a straight line when plotted on semilog paper. Curve 2 is termed the rectifier capacity as determined by the accelerated load limit tests.

3. A rating curve such as curve 3 is now determined by applying to the capacity curve a rating factor, which converts the values obtained on the accelerated load limit tests where arc back is imminent and will occur if load is increased 5% or 10% to rated values for operation in service with 20 years life and an arc-back rate of 2 to 3 arc backs per rectifier per year. This conversion factor is based on experience and relies heavily on the designer's judgment.

It should be recognized that the conversion factor is not a margin of safety. It is a conversion from values obtained on accelerated test to values obtained in service during the life of the machine.

Besides arc back, the conversion factor must take into account such long-term effects as metal fatigue, crystallization, corrosion, erosion, bombardment, coating of insulators, and other slow, long-term, physical processes which may lead to early failure. These factors have been evaluated for the double-wye connection. They must likewise be evaluated for the triple-diametric. The long life and high order of reliability established on rectifiers using the double-wye connection when operating at rating must be attained on the triple-diametric connection before its rating is valid. This does not mean that we must wait 8 to 10 years; rather, we must make an honest extrapolation of data obtained from 1- or 2-year field tests on equipment using the triple diametric circuit. Until this is done, the advantages in performance and economics that may be realized from this circuit

over those that have been demonstrated for the double-wye will not be known.

H. Winograd (Allis-Chalmers Manufacturing Company, Milwaukee, Wis.): The arc-drop voltage for the triple-diametric circuit was found to be nearly 1.5 volts lower than for the double-wye circuit. To what extent would this gain be offset by the additional losses in the larger rectifier and interphase transformers, as shown in Fig. 1, assuming the same d-c output?

It was stated in the paper that the lower arc-back current obtained with the triple-diametric circuit helps to overcome the problem of bracing the conductors of the rectifier transformer. Some calculations I made, using the data given in the paper, indicate that the forces in this transformer during an arc-back are higher.

The forces on a transformer leg are proportioned to the square of the ampere-turns. In the double-wye transformer, the phase connected to the faulted anode carries the arc-back current, while the opposite phase on the same leg has zero current. This is shown in Table II and was confirmed by other arc-back tests. In the triple-diametric transformer, the crest current in the opposite phase is about 80 per cent of the crest current in the faulted phase. Therefore, the sum of the ampere-turns of both phases has to be considered. In Fig. 5, the crest current points of the opposite phases 3 and 6 are displaced from each other. If the currents are added, the crest of their sum is about 85 per cent of the sum of their crests. A fair comparison of the forces can be made only on the basis of the same output rating for the two types of transformer connections, as given in Fig. 1.

Assuming that both transformers are designed with the same volts per turn, the voltages of the secondary phases can be used to represent turns in this analysis. With the values of fault currents and voltages given in Fig. 1, and the previously mentioned relationships of the opposite phase currents, the fault ampere-turns for the triple-diametric transformer are about 1.16 times those of the double-wye transformer. The ratio of the forces, proportional to the square of the ampere-turns, is about 1.35. The forces in the triple-diametric transformer during an arc-back are therefore about 35 per cent higher than in the double-wye transformer.

In considering the possible ratings rectifier tubes for the two circuits, the circuit duty on the tubes, as it affects arc-back probability, should be compared. For the same d-c output, as given in Fig. 1, the initial inverse and peak inverse voltages for the triple-diametric circuit are 50 per cent higher, while the anode current amplitude is two-thirds as high, their product being the same for the two circuits. In some of the literature on the theory of arc backs, the product of the initial inverse voltage and the rate of current change at the end of the commutating period was used as a criterion of the arc-back probability. On this basis the output current ratings for the two circuits of Fig. 1 should be the same. The use of such an arc-back factor is probably an oversimplification and cannot be applied for all types of tubes. Actual experience with rectifiers on the double-wye circuit indicates that when the output voltage is increased the current rating does not have to be reduced in the same proportion, and the kilowatt output can be increased. One possible reason for this is that the average value of the anode current is reduced as well as its amplitude. The average value of the anode current must also have an influence on the arc-back probability, because it affects the average mercury vapor pressure in the tube. In going from the double-wye to the triple-diametric circuit, the amplitude of the anode current and its rate of change at the end of the commutating period are reduced to two-thirds without reducing the average tube current. Perhaps the arc-back factor will have more significance here.

No amount of theorizing will settle this question of rectifier tube ratings. The ratings for the double-wye circuit were established on the basis of experience over a number of years in many installations. The same kind of experience is needed for rectifiers with the triple-diametric circuit. Have the authors operated tubes on a double-wye circuit, under similar service conditions, at tube currents approximating those mentioned in the paper, and if so, what was their relative performance?

The choice of a rectifier circuit is determined largely by economic considerations in the broader sense, taking into account such factors as cost, efficiency, rating, and performance. The decision is likely to be influenced by the type of service, the d-c output voltage, and the current and voltage capabilities of available rectifier tubes and switchgear.

The paper makes a valuable contribution.

R. V. Wachter, C. S. Hague, and **C. R. Marcum**: The discussions indicate a thorough study of the paper. The authors are grateful for the comments of men experienced in the rectifier field.

Mr. Winograd asked about the over-all effect on efficiency of the reduced arc losses and the increased transformer losses. On large units, the increased transformer losses offset about one-half of the gain in efficiency that is obtained from reduced arc voltage.

As for the operating results when tubes were operated at the same current on both the double-wye circuit and the triple-diametric circuit, during some field tests at the same average currents, other conditions were not the same. Laboratory tests at the same average current indicate that the tube performance on the doube-wye circuit is not as good as on the triple-diametric. Considering all the conditions under which these tests and others were made led to the conclusion that for some applications, the average tube can be appreciably increased without sacrifice in performance.

Mr. Herskind's discussion describes a system for rating tubes which, while different in detail, is basically the same as has been used by other rectifier manufacturers for many years. Experience with the detailed laboratory tests described by Mr. Herskind indicates that it results in an optimistic rating of the tubes.

Mr. Schmidt and Mr. Morton state that normal zig-zag interphase design might have about one-half the inductance indicated by the fault tests. If the interphase is properly designed to accomplish all the other desired performance characteristics of the triple-diametric circuit, its inductance is enough to limit the d-c feedback, as shown by the data in the paper.

To consider transformer reactance and its effect on fault current, the value of commutating reactance in ohms per phase gives the clearest indication of relative magnitude of fault current. For a given d-c regulation and a given output rating, the commutating ohms for the triple-diametric circuit are found from the fundamental voltage equation to be 0.239/0.106 times the commutating ohms for the double-wye circuit, neglecting the very minor differences introduced by IR drop or arc drop. Hence, the total fault current in the phase connected to the arcing-back tube is appreciably reduced in spite of the higher value of E_f. Expressed in per unit terms with transformer secondary rms ratings as the base, the ratio of triple-diametric reactance to double-wye reactance is approximately 1.4. We are unable to verify the ratio of per-unit reactance given by Mr. Schmidt and Mr. Morton; apparently it must be calculated on a different basis for the two circuits.

As Mr. Winograd points out, the vertical force on a given transformer leg is proportional to the square of the ampere-turns. However, when considering a through-winding relationship, consideration must be given to the mechanical displacement of the magnetic centers of each of the three coils in order to obtain the net vertical force on the transformer.

In the conventional double-wye transformer, the two secondary coils representing phases 1 and 4 are usually interwound in the form of a double-pitched helical coil. The primary winding is located concentrically around the two secondaries. This arrangement causes one secondary winding to have a slightly upward displacement and the other a slightly downward displacement with respect to the primary winding.

With the double-wye transformer under arc-back conditions, only phase 1 or phase 4 carries current. The net vertical force on the core leg is then proportional to the ampere-turns squared and multiplied by a constant proportional to the slight upward or downward displacement of the secondary winding that is carrying current.

In the case of the triple-diametric transformer, an arc-back produces nearly equal currents in both phase 1 and phase 4. Hence, the net displacement of the magnetic center of two secondary windings with respect to the primary winding approaches zero. Under this condition, the net vertical force on the leg is proportional to the sum of the ampere-turns squared and multiplied by the displacement constant, which is very small. The resultant net vertical force on the transformer leg is somewhat less than with a double-wye transformer.

All the discussers point out the importance of verifying through the medium of actual operating experience the gains in tube rating indicated by either theoretical considerations or limited laboratory data. We concur wholeheartedly with this concept, and it was for this reason that the experience was obtained which our paper records. This field experience has been so encouraging as to justify the application of the circuit with moderately increased tube ratings for several large installations. The practicality of further increases will be determined through experience.

Digital Compensation of Continuous-Data Feedback Control Systems

K. K. MAITRA
NONMEMBER AIEE

P. E. SARACHIK
STUDENT MEMBER AIEE

THE application of digital controllers or sampled-data processing units for stabilizing and compensating sampled-data feedback control systems has been a subject of considerable study in recent years.[1,2,3] It is now known that this technique is an effective and useful means in the design of sampled-data feedback systems.

This paper suggests some methods of employing digital computers most advantageously for compensating continuous-data feedback systems which function as control systems or regulators. The term "compensating" is used here to imply that no attempt is made to use the digital compensator to stabilize an inherently unstable system. It is only used to improve the dynamic performance of the system inasmuch as the rise time, maximum overshoot, or settling times are concerned. It will be seen later that, while the method of digital compensation presented here does not suffer any inherent limitations from the point of stabilizing unstable systems, certain specific advantages are achieved by restricting its application to compensation only.

In this paper two classes of systems are considered, namely duplicators and regulators. A duplicator is a control system which has either unity or a constant gain in the feedback line. Although duplicators are considered in this paper for the purpose of discussion, the methods of design developed here are equally applicable to the general class of control systems.

It is understood that the performance requirements of the two systems are entirely different. In the duplicator problem, the system is required to follow the reference input with no steady-state error, with the fastest possible rise time, and minimum overshoot. On the other hand, in the regulator system the output to any suddenly applied disturbance should decay rapidly and, if possible, with no overshoot. The methods of digital compensation described here can be used to optimize the performance of continuous systems.

Paper 56-210, recommended by the AIEE Feedback Control Systems Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Winter General Meeting, New York, N. Y., January 30-February 3, 1956. Manuscript submitted October 17, 1955; made available for printing December 13, 1955.

K. K. MAITRA and P. E. SARACHIK are with Columbia University, New York, N. Y.

The authors wish to thank Prof. Ragazzini for suggesting the topic of this paper, and for his continued advice, guidance, and encouragement. The research project was supported by the U. S. Air Force under *Contract No. AF 18(600)-677*, and monitored by the Office of Scientific Research, Air Research, and Development Command.

Description of Systems

The continuous-data feedback system shown in Fig. 1 is considered in all discussions of the duplicator problem. In the figure $D^*(s)$, the two switches, and the zero-order hold $H(s)$ represent the digital compensator unit.[1] The two switches shown are closed for a negligibly short time every T seconds. The hold system[4] clamps or hold the data pulse for the time interval between the sampling instants. The input $E_1^*(s)$ to $D^*(s)$ is in the form of discrete samples of the continuous error $E_1(s)$. The digital computing unit $D^*(s)$ stores a certain number of input pulses, does the same with a number of past samples of its own output, weights them in a specified way, and delivers a modified error pulse sequence $E_2^*(s)$ to the hold system. The output of the hold system $E_2(s)$ consists of a sequence of rectangular error pulses, the duration of each pulse being equal to the interval between two sampling instants.

The actuating error $E(s)$ to the plant is the sum of the continuous error $E_1(s)$ and the error pulse sequence $E_2(s)$. In the absence of the digital compensator, the only input to the plant is the continuous error $E_1(s)$ and the system reduces to a regular error-controlled unity-feedback continuous-data control system. Moreover, if the system without the compensator is such that it is normally stable and follows the input signal without any steady-state error, the digital compensating device will function only during the transient period and will have no effect whatsoever, on the system once the steady state condition has been reached. Clearly then, the compensating unit simply provides additional actuating data to improve the dynamic performance of the system. If however the plant $G(s)$ is such that the continuous system by itself needs a constant actuating error to follow the input signal with zero system error, the compensating device may be made to supply additional data at all times to force the system to follow the input. It is evident that in the latter case the output of the system will be calibrated by the performance of the compensating device. While this is not desirable, since the performance of the system is susceptible to any drift in the compensating device, this does not present a new problem in so far as the design method is concerned.

The regulator system considered is shown in Fig. 2, which can be further reduced to the equivalent system of Fig. 3. The operation of the compensating unit $D^*(s)$ in this system is the same as described in the previous section and does not require any further elaboration. As mentioned before, the regulator system is to be designed so that outputs due to inputs which usually appear as sudden disturbances will go to zero as rapidly as possible. Hence, the data sequence delivered by the compensating unit of a regulator performs a different function from that of a duplicator. The design problem is essentially that of choosing a suitable $D^*(s)$ in such a manner as to give a prescribed over-all performance to specified inputs.

Digital Compensator

The z-transform method can be readily used for the design of the systems under consideration.[1,4] It has been shown[1] that the z-transformed system function of a digital compensator is given by

$$D^*(z)=\frac{a_0+a_1z^{-1}+a_2z^{-2}+\ldots+a_lz^{-l}}{b_0+b_1z^{-1}+b_2z^{-2}+\ldots+b_kz^{-k}} \qquad (1)$$

The restriction on physical realizability is that no future samples but only present and past samples can be considered in realizing the system function of $D^*(s)$. Analytically this implies that, when $D^*(s)$ is expressed as a ratio of polynomials in z, the order of z in the numerator should not exceed that of the denominator. It can readily be seen that the constant b_0 in equation 1 must always be present to meet this condition.

The choice of $D^*(s)$ is not only limited by the physical realizability conditions but also is subject to certain restrictions which will be developed in the following sections.

Design of D*(z) for Duplicator Systems

It is shown in Appendix I, equation 53, that the output-input relation for the error-compensated system of Fig. 1 is

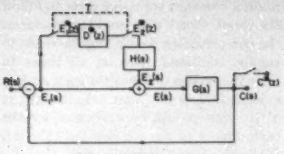


Fig. 1. Typical continuous-data control system using digital compensator

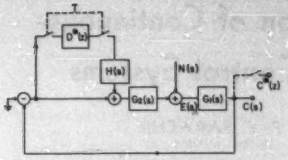


Fig. 2. Typical continuous-data regulating system using digital compensator

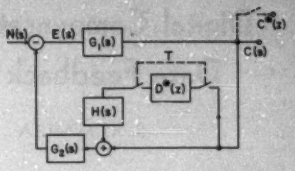


Fig. 3. System of Fig. 2 reduced

given by

$$C^*(z)=\frac{AR^*(z)+AH^*(z)D^*(z)R^*(z)}{1+AH^*(z)D^*(z)} \tag{2}$$

In equation 2, $AR^*(z)$ is the z transform of $A(s)\ R(s)$ in which $A(s)$ is the closed-loop transfer function of the continuous system without $D^*(z)$ inserted, that is

$$A(s)=\frac{G(s)}{1+G(s)}$$

and $R(s)$ is the Laplace transform of the input function. In the same way $AH^*(z)$ is the z transform of $A(s)\ H(s)$. If a new function $K_R{}^*(z)$ is now defined as the ratio of the output to the input, then from equation 2 it follows that

$$K_R{}^*(z)=\frac{C^*(z)}{R^*(z)}=\frac{\dfrac{AR^*(z)}{R^*(z)}+AH^*(z)D^*(z)}{1+AH^*(z)D^*(z)} \tag{3}$$

By rearranging equation 3

$$D^*(z)=\frac{1}{AH^*(z)}\,\frac{K_R{}^*(z)-\dfrac{AR^*(z)}{R^*(z)}}{1-K_R{}^*(z)} \tag{4}$$

It should be mentioned here that $K_R{}^*(z)$ can not be considered as the over-all system function because of its dependence on the input. Since each digital compensation program $D^*(z)$ shall be designed to yield a desirable over-all system response to a specific test input, it is quite legitimate to consider $K_R{}^*(z)$ as the true over-all transmission function for the specified input as a matter of convenience in the synthesis procedure.

Forms for $K_R{}^*(z)$

Limitations Due to Physical Realizability

The limitations imposed on the forms of $AH^*(z)$, $D^*(z)$, and $AR^*(z)/R^*(z)$ by their physical realizability restrict the choices of possible $K_R{}^*(z)$. By considering these limitations on the individual functions when expressed as ratios of polynomials in z^{-1}, it can be shown[1] that the expression for $K_R{}^*(z)$ has the form

$$K_R{}^*(z)=\frac{k_m z^{-m}+\ldots+k_p z^{-p}}{l_0+l_1 z^{-1}+\ldots+l_q z^{-q}} \tag{5}$$

where l_0 must always be present. For cases where there is no plant transfer lag, equation 5 becomes

$$K_R{}^*(z)=\frac{k_1 z^{-1}+k_2 z^{-2}+\ldots+k_p z^{-p}}{l_0+l_1 z^{-1}+\ldots+l_q z^{-q}} \tag{6}$$

Thus the classes of over-all response functions which can be obtained with the restrictions of physical realizability are those given by equations 5 and 6. However, there are still additional restrictions on the coefficients of $K_R{}^*(z)$ brought about by the fact that a prescribed performance is requested of the system in responding to a specificed test input. The usual functions used for this purpose are generally functions the Laplace transforms of which are of the form $1/s^m$, for $m \geqslant 1$. It is known[1,2] that the conditions, imposed by the requirement that the duplicator system follow such inputs with zero steady-state error, may be expressed by the relation

$$[1-K_R{}^*(z)]=(1-z^{-1})^m F_3{}^*(z) \tag{7}$$

where $F_3{}^*(z)$ is an undefined function with no zeros at $z=1$, and m is the order of the pole of $R(s)$, i.e., $1-K_R{}^*(z)$ must contain the factor $(1-z^{-1})^m$.

Limitations Due to Stability Requirements

It is seen from equations 2 or 3 that the stability of the over-all system is determined by the characteristic equation

$$1+AH^*(z)D^*(z)=0$$

This equation together with equation 3 indicates that, although the superficial gain function $K_R{}^*(z)$, as defined before, is dependent on the input signal, the stability of the system is not. This result is as should be expected in any physical system.

In order that the system may be stable, no zeros of the characteristic equation should lie on or outside the unit circle. In addition to this there are other conditions which will have to be satisfied for the system to be stable. The over-all transmission $K_R{}^*(z)$ as given by equation 3 is

$$K_R{}^*(z)=\frac{\dfrac{AR^*(z)}{R^*(z)}+AH^*(z)D^*(z)}{1+AH^*(z)D^*(z)}$$

It can be shown that, if $AH^*(z)$ has zeros at $|z|\geqslant 1$ and if $D^*(z)$ is so designed that it contains as its poles the undesirable zeros of $AH^*(z)$, stability of the system will depend on the exact cancellation of the zeros of $AH^*(z)$ and poles of $D^*(z)$ at $|z|\geqslant 1$. The validity of this statement is shown analytically in Appendix II, equation 56 and 57. It is also shown that zeros of $AH^*(z)$ on or outside the unit circle in the z-plane will appear as poles of $D^*(z)$ at $|z|\geqslant 1$ unless the function $K_R{}^*(z)-[AR^*(z)]/R^*(z)$ is made to contain the undesirable zeros of $AH^*(z)$. Since the stability of the system depends on the exact cancellation of the aforementioned poles and zeros, which is only ideally possible, the presence of undesirable zeros in $AH^*(z)$ imposes a restriction on the specification of $K_R{}^*(z)$. Stated more specifically, if $AH^*(z)=(1-nz^{-1})F^*(z)$, where $n\geqslant 1$ and $F^*(z)$ is a function of z^{-1} free of zeros at $|z|\geqslant 1$, then $K_R{}^*(z)-[AR^*(z)]/R^*(z)$ must be of the form

$$K_R{}^*(z)-\frac{AR^*(z)}{R^*(z)}=(1-nz^{-1})F_1{}^*(z) \tag{8}$$

where $F_1{}^*(z)$ is a function of z^{-1} with no zeros at $|z|\geqslant 1$.

The presence of zeros on or outside the unit circle of the function

$$P^*(z)-L^*(z)=0 \tag{9}$$

where $P^*(z)$ and $L^*(z)$ are related to $AR^*(z)/R^*(z)$ by

$$\frac{AR^*(z)}{R^*(z)}=\frac{L^*(z)}{P^*(z)} \tag{10}$$

also leads to instability because of improper cancellation. This difficulty can be avoided by containing the undesirable zeros of equation 9 in $1-K_R{}^*(z)$. The effect of the undesirable zeros of equation 9 on the stability of the over-all system is not intuitively obvious. It appears, however, to be true analytically, see Appendix II, equation 58.

It was indicated previously that the hold system associated with the compen-

sating unit is of the zero-order type, having the Laplace transform $H(s)=(1-e^{-sT})/s$. Therefore

$$AH^*(z)=Z[A(s)H(s)]=Z\left(A(s)\frac{1-e^{-sT}}{s}\right)$$

$$=(1-z^{-1})Z\left(\frac{A(s)}{s}\right) \quad (11)$$

In taking the z-transform of $[A(s)]/s$, it is found that one pole is introduced at $z=1$ because of the $1/s$ term of the hold. Since this pole is cancelled by the $(1-z^{-1})$ factor, all remaining poles of $AH^*(z)$ arise only from the poles of $A(s)$. For input functions having the form $R(s)=1/s^m$ and for $A(s)$ with no pole at $s=0$

$$AR^*(z)=Z[A(s)R(s)]=\frac{F_1^*(z)}{(1-z^{-1})^m} \quad (12)$$

where $F_1^*(z)$ is a function which is free of poles at $z=1$.

Thus

$$\frac{AR^*(z)}{R^*(z)}=F_1^*(z) \quad (13)$$

where $F_1^*(z)$ has no poles at $z=1$. Thus the poles of $[AR^*(z)]/[R^*(z)]$ are also those caused by the poles of $A(s)$. Hence the poles of $AH^*(z)$ and $[AR^*(z)]/[R^*(z)]$ are always identical, whereas this may or may not be true for the zeros.

In the special case where $R(s)=1/s$, $AH^*(z)$ and $[AR^*(z)]/[R^*(z)]$ are identical.

This property of automatic cancellation of the poles of $[AR^*(z)]/[R^*(z)]$ and $AH^*(z)$ is of great advantage. Examination of equation 4 shows that the poles of $AH^*(z)$ would appear as zeros of $D^*(z)$ if these were not cancelled by the poles of $[AR^*(z)]/[(R^*(z)]$. Cancellation of these poles makes it possible to impose no further restrictions on $K_R^*(z)$ because of the presence of unstable poles in $AH^*(z)$. This is not the case in most sample-data problems where special restrictions must be imposed on the over-all transmission function $K^*(z)$ to avoid instability which arises because of improper cancellations on or outside the unit circle.[2]

The limitations on the design of the digital compensating unit and on the specification of $K_R^*(z)$ are summarized in the following.

1. $K_R^*(z)$ must meet the physical realizability conditions expressed by equations 5 and 6.

2. In order that the system may follow specified test inputs of the form $R(s)=1/s^m$ without any error, the condition that $1-K_R^*(z)$ contains a factor $(1-z^{-1})^m$ must be satisfied.

3. The compensating unit $D^*(z)$ should not be designed to cancel zeros of the pulsed transfer function $AH^*(z)$ lying on or outside the unit circle. The undesirable zeros of $AH^*(z)$ must be contained in the function $K_R^*(z)-AR^*(z)/[R^*(z)]$ as expressed by equation 8.

4. The zeros of the function $[AR^*(z)]/[R^*(z)]-1=0$ at $|z|\geq 1$ must be contained in $1-K_R^*(z)$.

None of these restrictions offers any formal difficulty in the design procedure of the system. Observing them merely means that not all systems can be made to have the most desirable performance. Nevertheless, the performance of the compensated system is improved considerably.

Meeting all four conditions stated in the foregoing, $K_R^*(z)$ can always be chosen as the ratio of two polynomials, as expressed by equations 5 and 6. Evidently the transmission function will have an infinite settling time in much the same manner as the response of any linear system to a step disturbance. In digitally compensated continuous systems, as discussed here, it is also possible to specify $K_R^*(z)$ as a finite polynomial in z^{-1} of the form[1]

$$K_R^*(z)=k_m z^{-m}+\ldots+k_p z^{-p} \quad (14)$$

This will cause the system to have a finite settling time so far as the values at sampling instants are concerned. This may be a desirable objective, especially in the design of systems which are subjected to one particular type of test input. Evidently the finite polynomial form of $K_R^*(z)$ leads to the shortest possible error pulse sequence.

In certain design problems, it may be desirable to have a transmission function $K_R^*(z)$ with a specified denominator.[2,3] By following the general design procedure, it is possible to meet this requirement also.

Illustrative Example I

To illustrate the points brought out in the previous section, an example of the design of the program of a digital compensator will be given. The system to be designed is as shown in Fig. 1 with the following specifications

$$G(s)=\frac{1}{s(s+0.5)}$$

$$H(s)=\frac{1-e^{-sT}}{s}$$

$T=1.5$ seconds

$R(s)=1/s$

The system will be designed to be capable of following a unit step input function without error. Now

$$A(s)=\frac{1}{s^2+0.5s+1} \quad (15)$$

$A(s)$ is the over-all transfer function of the uncompensated continuous system. It will be observed from equation 15 that the uncompensated system is stable by itself and will follow a unit step input with zero steady-state error. The system is underdamped and the natural period of oscillation is found to be approximately 6.6 seconds. To improve the dynamic performance of this system, the sampling period T should be considerably smaller than the natural period of oscillation of the uncompensated system. In fact, the contribution of the digital compensator toward the improvement of the transient performance of the system will be more effective, the shorter the sampling period T. A very short T may, however, introduce considerable practical difficulty in the implementation of $D^*(z)$. While choice of any T is suitable for paper design, a compromise will have to be sought with due consideration ot the natural period of oscillation of the system and the implementation of $D^*(z)$. This justifies the choice of $T=1.5$ seconds which is approximately one quarter of the natural period of oscillation of the system. Then

$$AH(s)=(1-e^{-sT})\left(\frac{1}{s(s^2+0.5s+1)}\right) \quad (16)$$

Taking the z-transform of equation 16 for $T=1.5$ seconds yields

$$AH^*(z)=\left(\frac{a_1z^{-1}+a_2z^{-2}}{1-b_1z^{-1}+b_2z^{-2}}\right) \quad (17)$$

where $a_1=0.7423$, $a_2=0.5666$, $b_1=0.1635$, and $b_2=0.4724$. The values of a_1 and a_2 indicate that $AH^*(z)$ has no zero for $|z|\geq 1$.

The z-transforms can be conveniently taken by using the equation

$$G^*(z)=\frac{1}{2\pi j}\int_{c-j\infty}^{c+j\infty}\frac{G(s)}{1-e^{sT}z^{-1}}ds$$

Therefore $G^*(z)$ equals the sum of residues of $[G(s)]/(1-e^{sT}z^{-1})$ at the poles of $G(s)$ in the finite s plane.

It is interesting to note that $AH^*(z)|_{z=1}=1$. This result is in agreement with the fact that the continuous system follows a unit step input with zero steady-state error by itself.

Since $R(s)=1/s$, then $AR^*(z)/R^*(z)=AH^*(z)$. Since $AH^*(z)$ is free of zeros for $|z|\geq 1$, $K_R^*(z)$ can be chosen arbitrarily. Let

$$K_R^*(z)=k_1z^{-1}+k_2z^{-2} \quad (18)$$

It would be quite legitimate to choose $K_R^*(z)=z^{-1}$. It will, however, be realized that forcing an underdamped system to settle down too rapidly causes serious overshoots and ripple in the response between samples. Choice of $K_R^*(z)$, as

in equation 18, is therefore a reasonable compromise.

The condition set forth in equation 7 must also be satisfied for the system to follow the input signal. Thus

$$1-K_R^*(z)=(1-z^{-1})(1+l_1z^{-1}) \tag{19}$$

From equations 18 and 19

$$1-k_1z^{-1}-k_2z^{-2}=1+(l_1-1)z^{-1}-l_1z^{-2} \tag{20}$$

$$\therefore k_1=1-l_1 \tag{21}$$

$$k_2=l_1 \tag{22}$$

From equations 21 and 22

$$k_1+k_2=1 \tag{23}$$

Substituting equations 17, 18, and 19 in equation 4 and simplifying gives

$$D^*(z)=\frac{k_1(1-b_1z^{-1}+b_2z^{-2})\left(1+\frac{k_2}{k_1}z^{-1}\right)-a_1^*\left(1+\frac{a_2}{a_1}z^{-1}\right)}{a_1\left(1+\frac{a_2}{a_1}z^{-1}\right)(1-z^{-1})(1+l_1z^{-1})} \tag{24}$$

By choosing $k_2/k_1=a_2/a_1$, equation 24 further simplifies to

$$D^*(z)=\frac{\frac{k_1}{a_1}(1-b_1z^{-1}+b_2z^{-2})-1}{(1-z^{-1})(1+l_1z^{-1})} \tag{25}$$

also

$$k_1=\frac{a_1}{a_1+a_2}$$

$$l_1=\frac{a_2}{a_1+a_2}$$

Substituting for k_1 and l_1 in equation 25 and using the property that $(a_1+a_2)/(1-b_1-b_2)=1$, gives

$$D^*(z)=\frac{\frac{1}{a_1+a_2}(1-a_1-a_2-b_2z^{-1})}{1+\frac{a_2}{a_1+a_2}z^{-1}} \tag{26}$$

Finally substituting for a_1, a_2, and b_2

$$D^*(z)=\frac{-0.236-0.361z^{-1}}{1+0.433z^{-1}} \tag{27}$$

also

$$K_R^*(z)=0.567z^{-1}+0.433z^{-2} \tag{28}$$

It is evident from equation 27 that the digital computer unit must store one past input sample and one past sample of its own output. This is a relatively simple form for $D^*(z)$. Now

$$C^*(z)=K_R^*(z)R^*(z) \tag{29}$$

Substituting the z transform of the unit step input $1/(1-z^{-1})$ into equation 29, the output sequence is obtained by long division. Thus

$$C^*(z)=0.567z^{-1}+z^{-2}+z^{-3}+\ldots \tag{30}$$

For the purposes of comparison, the uncompensated response $C_u^*(z)$ of this system at sampling instants is obtained from

$$C_u^*(z)=AR^*(z)$$

In this example

$$R^*(z)=\frac{1}{1-z^{-1}}\text{ (unit step)}$$

$$C_u^*(z)=\frac{0.742z^{-1}+0.566z^{-2}}{(1-z^{-1})(1-0.164z^{-1}+0.472z^{-2})}=$$

$$0.742z^{-1}+1.430z^{-2}+1.193z^{-3}+0.830z^{-4}+\ldots$$

It is apparent that the transient performance of this system is greatly improved by digital compensation. It might also be noted that, since the uncompensated system follows a unit step with zero steady-state error, the digital computer will supply no program to the system under steady-state conditions.

The object of the design procedure discussed in the previous section was to synthesize a program for a digital compensator unit which would yield the best possible over-all transmission function $K_R^*(z)$. In the illustrative example given, it was found that $D^*(z)$ has a simple form. In many systems a more complex program may be required to obtain a specified $K_R^*(z)$. Now an alternative method of design will be discussed. The principal idea of this method lies in the fact that a very simple form of $D^*(z)$ with arbitrary coefficients is specified. The object then is to adjust these coefficients until a satisfactory transmission function $K_R^*(z)$ for the specified test input is obtained. This method can be illustrated by considering a problem in general form. The system considered is as shown in Fig. 1. For simplicity the system will be designed to follow a unit step input.

Taking a physical system having $AH^*(z)$ of the form

$$AH^*(z)=\frac{a_1z^{-1}+a_2z^{-2}}{1+b_1z^{-1}+b_2z^{-2}} \tag{31}$$

The input is a unit step function and, therefore, $[AR^*(z)]/[R^*(z)]=AH^*(z)$. Let $D^*(z)$ be specified as

$$D^*(z)=\frac{c_0+c_1z^{-1}}{1+d_1z^{-1}} \tag{32}$$

Substituting equations 31 and 32 in equation 3

$$K_R^*(z)=\frac{[1+c_0+(c_1+d_1)z^{-1}](a_1z^{-1}+a_2z^{-2})}{1+(d_1+a_1c_0+b_1)z^{-1}+(b_2+b_1d_1+c_0a_2+c_1a_1)z^{-2}+(b_2d_1+a_2c_1)z^{-3}} \tag{33}$$

By adjusting the unknown constants c_0, c_1, and d_1 in equation 32, it is possible to synthesize a satisfactory $K_R^*(z)$ within certain limits. Clearly, there would be more flexibility on the choice of a satisfactory $K_R^*(z)$ if a more complicated $D^*(z)$ had been assumed to start with; i.e., a $D^*(z)$ with a greater number of storages and, therefore, more arbitrary coefficients. For the relatively simple case of equation 32, there are enough arbitrary coefficients so that a satisfactory $K_R^*(z)$ may be specified. What would be the most desirable form for $K_R^*(z)$ is an eternal problem in the design of servo systems. This question is usually decided on the basis of the requirements of individual systems. Let it be assumed in this case that $K_R^*(z)$ is desired to have a finite settling time. This condition can be obtained by equating the three coefficients in the denominator polynomial to zero and solving for the unknown constants c_0, c_1, and d_1.

If in an alternative approach it is desired that the denominator be a polynomial in z^{-1}, such as $1+b_1z^{-1}+b_2z^{-2}+b_3z^{-3}$, then the condition to be satisfied for stability of $K_R^*(z)$ is that the characteristic equation

$$1+b_1z^{-1}+b_2z^{-2}+b_3z^{-3}=0$$

must not have any zero on or outside the unit circle. This can be easily done by using any one of the usual stability criteria to obtain the coefficients c_0, c_1, and d_1.[4] It is apparent from equation 33 that this method automatically takes care of the condition for stability discussed previously and expressed by equation 8. The second condition for stability as given in equation 62, Appendix II, is also automatically satisfied.

If it is assumed that the uncompensated continuous system follows the input test signal by itself, then, as pointed out previously, $AH^*(z)|_{z=1}=1$. By using this property of $AH^*(z)$, it can be easily verified that the function $1-K_R^*(z)$ obtained by the method of this section always contains a factor $(1-z^{-1})$ and thus satisfies the necessary condition expressed by equation 7.

Design of D*(z) for Regulator System

The method of designing programs for a digital compensator that will yield satisfactory performance for the regulator system of Fig. 2 follows the same line as that of the duplicator system discussed in the previous sections. It is shown in Appendix III, equation 69, that the input-output relation is given by

$$C^*(z)=\frac{PN^*(z)}{1+PHG_3^*(z)D^*(z)}$$

where

$PN^*(z)=z$ transform of $P(s)N(s)$
$N(s)=$disturbance input
$PHG_3^*(z)=z$ transform of $P(s)H(s)G_3(s)$

$$P(s)=\frac{G_1(s)}{1+G_1(s)G_2(s)}$$

If a new function $K_N^*(z)$ is defined as the ratio of the output to the input, then

$$K_N^*(z)=\frac{\frac{PN^*(z)}{N^*(z)}}{1+PHG_3^*(z)D^*(z)} \tag{34}$$

By rearranging equation 34

$$D^*(z)=\frac{1}{PHG_3^*(z)}\,\frac{\frac{PN^*(z)}{N^*(z)}-K_N^*(z)}{K_N^*(z)} \tag{35}$$

Forms for $K_N^*(z)$

1. The limitations imposed on the forms of $K_N^*(z)$ because of the requirements of physical realizability are identical to those discussed in the duplicator problem and, hence, equations 5 and 6 also apply to this case.

2. In a regulator system, the performance of the system should be such that the outputs to specific test input disturbances die out to zero as fast as possible. The functions usually used for this purpose are those whose Laplace transforms are of the form $1/s^m$ for $m\geq 1$. By application of the final value theorem it can be shown that the condition imposed by the foregoing requirement may be satisfied by the relation

$$K_N^*(z)=(1-z^{-1})^m K_{1N}^*(z) \tag{36}$$

where $K_{1N}^*(z)$ is a function free of zeros at $z=1$.

3. The compensating unit $D^*(z)$ should not be designed to cancel zeros of the pulsed transfer function $PHG_3^*(z)$ lying on or outside the unit circle. The undesirable zeros of $PHG_3^*(z)$ must be contained in the function $[PN^*(z)]/[N^*(z)]-K_N^*(z)$, i.e.

$$\frac{PN^*(z)}{N^*(z)}-K_N^*(z)=(1-az^{-1}) \tag{37}$$

where $(1-az^{-1})$ is the zero of $PHG_3^*(z)$ at $|z|\geq 1$.

4. The zeros of the function $[PN^*(z)]/[N^*(z)]$ at $|z|\geq 1$ must be contained in $K_N^*(z)$.

5. If $(1-bz^{-1})$ is a common zero of both $PHG_3^*(z)$, and $[PN^*(z)]/[N^*(z)]$ at $z=b\geq 1$, then, to avoid instability caused by improper cancellation

$$\frac{PN^*(z)}{N^*(z)}-K_N^*(z)=(1-bz^{-1})^3F_5^*(z) \tag{38}$$

in which $F_5^*(z)$ has no zeros or poles on or outside the unit circle. This condition always arises when $N^*(z)$ is a unit step disturbance and $PHG_3^*(z)$ has zeros on or outside the unit circle.

That the foregoing conditions are necessary and sufficient to ensure stability of the over-all system can be verified in a manner similar to that shown in Appendix II for the duplicator problem.

As was mentioned previously, the object of employing digital controllers is simply to improve the transient performance of the system. As such, the only systems considered in this discussion are those which are normally stable by themselves. If this is true, then neither $PHG_3^*(z)$ nor $[PN^*(z)]/[N^*(z)]$ can have any poles on or outside the unit circle. Therefore, the consideration of unstable poles of these functions do not enter into the present problem.

In the regulator problem also, it is always possible to specify $K_N^*(z)$ as a finite polynomial in the form of equation 14. This implies that the output would be reduced to zero in a finite number of sampling instants. However, in some cases, specifying a finite polynomial form of $K_N^*(z)$ may require considerable complexity of $D^*(z)$. Simplicity of $D^*(z)$ is always a desirable factor in the synthesis of digitally controlled systems and, therefore, it may be more desirable in some cases to specify a $K_N^*(z)$ as a ratio of two polynomials in order to achieve this objective.

Illustrative Example II

To illustrate the method of designing a suitable program for a digital controller which will yield satisfactory over-all performance of the regulator system a simple illustrative example will be given. The system under consideration is as shown in Fig. 3.

Let $G_2(s)=K_2$, which is a constant gain, and

$$G_1(s)=\frac{1}{s(s+0.5)}$$

then

$$P(s)=\frac{1}{s^2+0.5s+K_2} \tag{39}$$

Equation 39 is the over-all transfer function of the continuous system. This equation also shows that the continuous system is stable by itself for all K_2. A value of $K_2=10$, which is typical for a regulator system, is chosen for this example. Now

$$N(s)=\frac{1}{s} \tag{40}$$

Without digital compensation, the output of the continuous system itself, because of the sudden unit step disturbance, is given by

$$c(t)=\frac{1}{10}\left[u(t)-e^{-0.25t}\left(\cos 3.19t-\frac{0.25}{3.19}\times\sin 3.19t\right)u(t)\right] \tag{41}$$

This shows that the final value of the output will be 1/10 when the input disturbance is a unit step function of unit amplitude. It will now be shown that, by introducing $D^*(z)$, the output caused by this disturbance can be reduced to zero in a finite time. A value of 1 second is chosen for T.

Now

$$P(s)H(s)G_2(s)=\frac{10(1-e^{-sT})}{s(s^2+0.5s+10)}$$

Taking the z transform, the pulsed transfer function becomes

$$PHG_3^*(z)=\frac{1.786z^{-1}+1.378z^{-2}}{1+1.558z^{-1}+0.606z^{-2}} \tag{42}$$

Also

$$\frac{PN^*(z)}{N^*(z)}=\frac{1}{10}PHG_3^*(z)$$

In accordance with the condition stated by equation 36, let

$$K_N^*(z)=(1-z^{-1})(a_1z^{-1}) \tag{43}$$

Then, by equation 35

$$D^*(z)=\frac{0.1786+0.1378z^{-1}-a_1(1-z^{-1})}{(1+1.558z^{-1}+0.606z^{-2})}\,\frac{(1+1.558z^{-1}+0.606z^{-2})}{a_1(1-z^{-1})(1.786z^{-1}+1.378z^{-2})} \tag{44}$$

As mentioned previously, in order that $D^*(z)$ may be physically realizable, the constant b_0 in equation 1 must be present. It is seen in equation 44 that for physical realizability of $D^*(z)$ the constant a_1 must be equal to 0.1786. Thus

$$D^*(z)=\frac{0.0384+0.17z^{-1}+0.108z^{-2}}{(1.786+1.378z^{-1})(1-z^{-1})} \tag{45}$$

With this $D^*(z)$, the system transmission function is

$$K_N^*(z)=(1-z^{-1})0.1786z^{-1}$$

and therefore the output resulting from the test input disturbance function $1/1-z^{-1}$ will simply be

$$C_N^*(z)=0.1786z^{-1}$$

which shows that the output reduces to zero at the second sampling instant.

It is seen that the coefficients of the numerator polynomial of equation 45 are very small because a value of $K_2=10$ was assumed. Realization of these small weighting factors in a digital controller may introduce considerable difficulty. To avoid this difficulty it may be necessary to reduce the value of the constant gain K_2.

Alternative Method of Design

The alternative method of design suggested for the duplicator system may also be conveniently applied to the synthesis of the regulator system. The steps involved in the design procedure are essentially the same. There is, however, one important point that should be brought out to make the alternative method complete.

It is known that the output of a regulator system to a suddenly applied input would normally reduce to zero in the steady state if the feedback loop contains integrating elements of sufficient order. This is valid under the assumption that the input function has a Laplace transform of the form $1/s^m$ for $m \geqq 1$. In other words, the number of integrations in the feedback loop must be at least as high as the order of s in the equation $N(s)=1/s^m$ if the system is to reduce the output to zero in the steady state. If the feedback loop of the continuous regulator system has integrating elements of the required order, then the function of the digital compensator is simply to provide additional data to force the output to reach zero as fast as possible.

It is therefore evident that, in the absence of enough integrating elements in the feedback loop of the continuous regulator system, the digital controller $D^*(z)$ must contain integrations of the required order to force the output to zero. In the example of the previous section this requirement is satisfied by the appearance of the factor $(1-z^{-1})$ in the denominator. This factor is nothing but the z transform of a first-order integration $1/s$.

In the light of the foregoing discussion it is now possible to set a condition in the alternative method of design. In assuming $D^*(z)$ arbitrarily, a factor $(1-z^{-1})^{m-n}$ must be included in the denominator of $D^*(z)$, where n is the order of integrations present in $G_2(s)$ in the feedback loop of the system of Fig. 3 and m is the order of s in the input function $N(s)=1/s^m$. With this restriction on the form of $D^*(z)$ the system can be designed in exactly the same manner as that which was discussed previously.

It should, however, be noted that it is not desirable to design the compensating unit to contain more than one integration; see discussion of reference 1 by Salzer. This imposes a limitation on the applicability of this method, particularly where the input disturbance function is a ramp and there is no integrating element present in the feedback loop of the original system.

Effect of Delayed Input

The preceding analysis was based on the tacit assumption that the input to the system is applied at the instant of sampling. While this assumption leads to relatively simple analytical work, in physical systems this is seldom valid; i.e., the instant at which the input time function is applied may correspond to any value of time during the interval between two sampling instants. This results in the continuous system beginning to respond before the digital compensator becomes aware of the presence of the input. Therefore, during the short period of time from the instant of application of the input to the next sampling instant the system behaves as though the compensator were not present and, as such, the output of the system cannot be controlled at the first sampling instant. If this interval of time is denoted by ΔT, where T is the sampling period and Δ is a number between 0 and 1, the output at the first sampling instant will be a function of Δ as well as of T. The factor Δ is a random variable and completely beyond the control of the designer. This results in a system which is essentially time-variant, in the sense that the output will be dependent on the moment at which the input is applied.

Analysis of the system under this condition is not particularly difficult since the effect may be considered merely by applying an input of the form $R(s,\lambda)=R(s)\epsilon^{-\lambda Ts}$ to the system previously discussed. The delay factor is related to Δ by the relation $\lambda+\Delta=1$ and hence its value also lies between zero and one. The z transform of functions of the form $F(s,\lambda) = F(s)\epsilon^{-\lambda Ts}$ may be taken by methods described in reference 1, and this results in

$$Z[F(s,\lambda)] = z^{-1}Z[F(s)\epsilon^{\Delta Ts}] \triangleq z^{-1}F^*(z,\Delta)$$

It can be seen that the over-all transmission function $K_R^*(z)$, in this case, will be a function of the delay factor λ. The designer can only specify a response for one particular value of λ. The result is that a system designed with this specification will yield different responses for inputs that have been applied at different instants.

In view of the nature of this problem it is difficult to suggest a definite design criterion that will take care of the delay factor. One way of reducing the effect of the delay on the performance of the system is to increase the rate of sampling. The system will then have less time to "run free" after the application of the input signal.

Even with a fast rate of sampling the effect of the delay factor λ will not be completely eliminated. Because λ is a random variable which may be described by a uniform probability distribution, the specification of the over-all transmission function $K_R^*(z)$ could be made on the basis of the average value of λ which is 0.5. This is of dubious advantage however, and less analytical work is involved if no delay is assumed. If the sampling rate is sufficiently fast, the results will be satisfactory.

In the case of the duplicator system, the effect of delay is often much more pronounced than it is in the regulator. This is because in the duplicator the assumption of no delay implies that the value of the output at the first sampling instant $t=T$ can be arbitrarily specified. If the input is delayed, the value of the output at the first sampling instant after application of the input cannot be arbitrarily specified. This value is determined only by the response of the uncompensated system and the instant at which the input is applied. This restriction is especially important for underdamped systems, since too long a sampling period might give the system a chance to overshoot considerably before any correction is applied by the compensator.

In the regulator system, whenever the plant $G_1(s)$ is of low pass, the application of any nonimpulsive input will produce zero output at the moment of application. When the input is applied at a sampling instant $t=0$, the digital compensator operating on the output does not begin to deliver its program until the following sample $t=T$. The result is that the earliest instant at which the value of the output can be arbitrarily specified is $t=2T$. If a delayed input is now considered, the situation is essentially the same, i.e., for an input applied at $t=\lambda T$, the compensator delivers its first pulse at $t=T$ and, hence, the value of the output at $t=2T$ is still the first one that is under the control of the designer.

Response curves to step inputs, obtained from analogue-computer studies of the control system described in the section entitled "Illustrative Example I," are presented in Fig. 4. Fig. 4(A) shows the response of the uncompensated system. In Fig. 4(B) and (C) the variations of the response caused by different values of the delay factor λ can be seen.

Fig. 5 shows the responses of the same continuous system but with a program of $D^*(z) = (1.22 - 1.525z^{-1})/(1-0.469z^{-1})$ and a sampling period $T=0.75$. This program of $D^*(z)$ was obtained by the method described in the section entitled

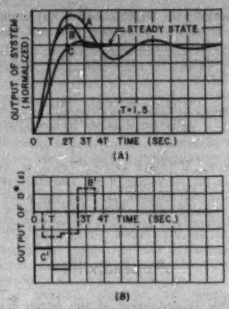


Fig. 4. Experimental response curves of system of Fig. 1 used in illustrative example I with T=1.5 seconds

A—Response of uncompensated system used in illustrative example I
B—Response of compensated system showing effect of delayed input
C—Response of compensated system when input appears at the instant of sampling
O—Calculated response of compensated systems at sampling instants
B'—Output pulse sequences of the digital compensator corresponding to the response curve B
C'—Output pulse sequences of the digital compensator corresponding to the response curve C

"Illustrative Example I." The response curves show that the effect of the delay in causing variations of the over-all response is less pronounced when a faster sampling rate is used.

Conclusion

Systems involving plant transfer lags and problems of plant saturation are not included in this paper. These additional considerations do not, however, present any formal difficulty in the synthesis procedure discussed here. Detailed discussions of these points appear in reference 1 and the same principles can be easily extended to the present problem. Problems relating to the implementation of the digital compensator unit and details of operation are beyond the scope of this paper.

An alternative configuration for compensation of control systems, as shown in Fig. 6, appears to be a possibility. This system is identical to the regulator system of Fig. 3, when $G_2(s)=1$, although the performance requirements are different. Because of the appearance of the compensating unit in the feedback loop, the output is calibrated even when the system has arrived at the steady state. Thus,

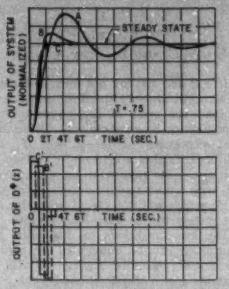


Fig. 5. Experimental response curves of system of Fig. 1 using program with T=0.75 second

See Fig. 4 for definitions

steady-state performance of the system is extremely susceptible to any variations in the coefficients of the digital controller. This problem does not arise when the compensating unit is introduced in the feed-forward loop. Thus, this scheme is not very desirable from a practical viewpoint.

Methods of compensating continuous-data feedback systems by employing a digital controller have been suggested. The discussions cover a thorough treatment of the methods of synthesizing control and regulating systems. The primary object of this type of compensation is to improve the dynamic performance of stable continuous systems.

No attempt has been made to use the digital compensator to stabilize an unstable system. This means that, in the event of failure of the digital controller unit, the system will still function although not with the same quality of transient performance. This is undoubtedly a very desirable feature of this type of compensation.

Appendix I.

Referring to Fig. 1, it is seen that

$$C(s)=G(s)E(s) \tag{46}$$

$$E(s)=E_1(s)+E_1^*(s)D^*(s)H(s) \tag{47}$$

and substituting into equation 46 gives

$$C(s)=G(s)E_1(s)+G(s)H(s)D^*(s)E_1^*(s) \tag{48}$$

Now it is easily seen that

$$E_1(s)=R(s)-C(s)$$

and taking the z transform

$$E_1^*(s)=R^*(s)-C^*(s)$$

Substituting these equations into equation 48 gives

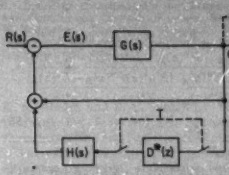


Fig. 6. Continuous-data control system using a digital compensator across the feedback loop

$$C(s)=G(s)R(s)-G(s)C(s)+G(s)H(s)D^*(s)R^*(s)-G(s)H(s)D^*(s)C^*(s) \tag{49}$$

Collecting terms

$$C(s)=\frac{G(s)}{1+G(s)}R(s)+\frac{G(s)}{1+G(s)}H(s)D^*(s)-R^*(s)-\frac{G(s)}{1+G(s)}H(s)D^*(s)C^*(s) \tag{50}$$

Let

$$A(s)=\frac{G(s)}{1+G(s)}$$

Therefore

$$C(s)=A(s)R(s)+A(s)H(s)D^*(s)R^*(s)-A(s)H(s)D^*(s)C^*(s) \tag{51}$$

Taking the z transform

$$C^*(s)=AR^*(s)+AH^*(s)D^*(s)R^*(s)-AH^*(s)D^*(s)C^*(s) \tag{52}$$

Rearranging and collecting terms

$$C^*(s)=\frac{AR^*(s)+AH^*(s)D^*(s)R^*(s)}{1+AH^*(s)D^*(s)} \tag{53}$$

Appendix II.

The stability conditions for $K_R^*(s)$ are to be shown: Consider $AH^*(s)$ of the form

$$AH^*(s)=\frac{N^*(s)}{P^*(s)}$$

where

$N^*(s)$ = the numerator polynomial in z^{-1}
$P^*(s)$ = the denominator polynomial in z^{-1}

Now since the poles of $AH^*(s)$ and $[AR^*(s)]/[R^*(s)]$ arise only from the poles of $A(s)$, $[AR^*(s)]/[R^*(s)]$ will be of the form

$$\frac{AR^*(s)}{R^*(s)}=\frac{L^*(s)}{P^*(s)}$$

Substituting these expressions into equation 4 for $D^*(s)$

$$D^*(s)=\frac{P^*(s)}{N^*(s)}\frac{K_D^*(s)-\frac{L^*(s)}{P^*(s)}}{1-K_D^*(s)}=\frac{K_D^*(s)P^*(s)-L^*(s)}{N^*(s)[1-K_D^*(s)]} \tag{54}$$

where $K_D^*(s)$ denotes the desired transmission function. Substituting this expression into equation 3 for $K_R^*(s)$

$$K_R^*(z)=\frac{\dfrac{L^*(z)_A}{P^*(z)_A}+\dfrac{N^*(z)_A}{P^*(z)_A}\dfrac{K_D^*(z)P^*(z)_D-L^*(z)_D}{N^*(z)_D[1-K_D^*(z)]}}{1+\dfrac{N^*(z)_A}{P^*(z)_A}\dfrac{[K_D^*(z)P^*(z)_D-L^*(z)_D]}{N^*(z)_D[1-K_D^*(z)]}} \tag{55}$$

$$K_R^*(z)=\frac{L^*(z)_A N^*(z)_A[1-K_D^*(z)]+N^*(z)_A\times[K_D^*(z)P^*(z)_D-L^*(z)_D]}{P^*(z)_A N^*(z)_D[1-K_D^*(z)]+N^*(z)_A\times[K_D^*(z)P^*(z)_D-L^*(z)_D]}$$

where the subscript A indicates that the zeros of the associated polynomial are caused by the system $A(s)$, while the subscript D indicates zeros caused by the settings in $D^*(z)$.

If the coefficients of $D^*(z)$ were set perfectly so that $L^*(z)_D=L^*(z)_A$, $P^*(z)_D=P^*(z)_A$ and $N^*(z)_D=N^*(z)_A$, it is easily seen that equation 55 reduces to

$$K_R^*(z)=K_D^*(z)$$

which shows that the actual transmission function is the one desired.

Consider now the case where $AH^*(z)$ has a zero outside the unit circle. This zero appears in the numerator polynomial $N^*(z)$, which for simplicity will be taken just as $(1-nz^{-1})_A$ where $n>1$; then

$$K_R^*(z)=\frac{L^*(z)_A(1-nz^{-1})_D[1-K_D^*(z)]+(1-nz^{-1})_A[K_D^*(z)P^*(z)_D-L^*(z)_D]}{(1-nz^{-1})_D P^*(z)_A[1-K_D^*(z)]+(1-nz^{-1})[K_D^*(z)P^*(z)_D-L^*(z)_D]} \tag{56}$$

Note that, if $(1-nz^{-1})_D=(1-nz^{-1})_A$, then this factor may be divided out. If, on the other hand these factors differ slightly because of unprecise setting of, or drift in, the coefficients of $D^*(z)$, then the numerator of equation 56 will have a zero at some $z=n_T$ such that n_T lies between n_A and n_D; the denominator will also have a zero at $z=n_B$ such that n_B lies between n_A and n_D. Note that $n_B\neq n_T$ since the numerator and denominator functions are different.

Therefore $K_R^*(z)$ can be written

$$K_R^*(z)=\frac{(1-n_T z^{-1})}{(1-n_B z^{-1})}\times\frac{K_D^*(z)[P^*(z)_D-L^*(z)_A]+[L^*(z)_A-L^*(z)_D]}{[P^*(z)_A-L^*(z)_D]+K_D^*(z)[P^*(z)_D-P^*(z)_A]} \tag{57}$$

It has been shown[1,2] that the imperfect cancellation of $(1-n_B z^{-1})$ by $(1-n_T z^{-1})$ outside the unit circle leads to instability. Now let

$$L^*(z)_A-L^*(z)_D=\epsilon(z)$$

and

$$P^*(z)_A-P^*(z)_D=\delta(z)$$

then

$$K_R^*(z)=K_D^*(z)\frac{(1-n_T z^{-1})}{(1-n_B z^{-1})}\times\frac{P^*(z)_A-L^*(z)_A+\dfrac{1}{K^*(z)}\epsilon(z)-\delta(z)}{P^*(z)_A-L^*(z)_A+\epsilon(z)-K_D^*(z)\delta(z)} \tag{58}$$

If $\epsilon(s)$ and $\delta(s)$ are not identically equal to zero but are very small for all values of s, then the zeros of the numerator and denominator lie very close to the zeros of the expression $P^*(z)-L^*(z)$. If now $P^*(z)-L^*(z)$ has a zero on or outside the unit circle, the numerator and denominator will each have a zero on or outside the unit circle which will not cancel perfectly, and the system will be unstable as before.

The two conditions of instability just mentioned can be eliminated by making the proper choice of $K_D^*(z)$. If $AH^*(z)$ contains a zero outside the unit circle

$$AH^*(z)=\frac{(1-nz^{-1})}{P^*(z)} \tag{59}$$

then make $K^*(z)-AR^*(z)/R^*(z)$ contain this zero as a factor, i.e.

$$K^*(z)-\frac{AR^*(z)}{R^*(z)}=\frac{(1-nz^{-1})F^*(z)}{P^*(z)}$$

Substituting equation 59 into equation 4 for $D^*(z)$ and canceling appropriate terms, the resulting expression for $D^*(z)$ is

$$D^*(z)=\frac{F^*(z)}{1-K_D^*(z)} \tag{60}$$

Substituting equation 60 into equation 3 for $K_R^*(z)$ gives

$$K_R^*(z)=\frac{L^*(z)_A[1-K^*(z)]+(1-nz^{-1})_A F^*(z)_D}{P_A^*(z)+(1-nz^{-1})_A F^*(z)_D} \tag{61}$$

Note that neither numerator nor denominator has a zero near $z=n$. Therefore instability caused by this condition has been eliminated.

If $P^*(z)-L^*(z)$ contains a zero on or outside the unit circle

$$P^*(z)-L^*(z)=(1-qz^{-1})Q^*(z)$$

where $q\geq 1$ and $Q^*(z)$ has no zeros on or outside the unit circle. Also let $AH^*(z)=[N^*(z)]/P^*(z)]$, where $N^*(z)$ has no zeros on or outside the unit circle. Then

$$D^*(z)=\frac{K_D^*(z)P^*(z)-L^*(z)}{[1-K_D^*(z)]N^*(z)}=\frac{K_D^*(z)P^*(z)+(1-qz^{-1})Q^*(z)-P^*(z)}{[1-K_D^*(z)]N^*(z)}=\frac{(1-qz^{-1})Q^*(z)}{[1-K_D^*(z)]N^*(z)}-\frac{P^*(z)}{N^*(z)} \tag{62}$$

Now make $1-K_D^*(z)$ contain the factor $(1-qz^{-1})$

$$1-K_D^*(z)=(1-qz^{-1})M^*(z) \tag{63}$$

where $M^*(z)$ has no poles outside the unit circle

$$D^*(z)=\frac{Q^*(z)-P^*(z)M^*(z)}{N^*(z)M^*(z)} \tag{64}$$

When equation 63 for $D^*(z)$ is substituted into equation 3 for $K_R^*(z)$ and $N_A^*(z)$ is canceled with $N_D^*(z)$, since $N^*(z)$ contains no zeros on or outside the unit circle, imperfect cancellation will not result in instability; the resulting expression for $K_R^*(z)$ is

$$K_R^*(z)=\frac{Q^*(z)_D-M^*(z)_D[P^*(z)_A-L^*(z)_A]}{Q^*(z)_D-M^*(z)_D[P^*(z)_A-P^*(z)_D]} \tag{65}$$

Now $M^*(z)_D[P^*(z)_A-P^*(z)_D]$ is very small for all values of z. Therefore zeros of the denominator lie near the zeros of $Q^*(z)_D$ and, since $Q^*(z)_D$ has no zeros on or outside the unit circle, $K_R^*(z)$ is stable.

Appendix III.

This is derivation of the output-input relation of the regulator system in Fig. 3.

$$C(s)=E(s)G_1(s) \tag{66}$$

$$E(s)=N(s)-[C(s)+C^*(s)D^*(s)H(s)]G_2(s) \tag{67}$$

Substituting for $E(s)$ in equation 66

$$C(s)=N(s)G_1(s)-G_1(s)G_2(s)C(s)-C^*(s)D^*(s)G_1(s)G_2(s)H(s) \tag{68}$$

Rearranging equation 68

$$C(s)=\frac{N(s)G_1(s)}{1+G_1(s)G_2(s)}-\frac{G_1(s)}{1+G_1(s)G_2(s)}\times G_2(s)H(s)C^*(s)D^*(s)$$

Let

$$\frac{G_1(s)}{1+G_1(s)G_2(s)}=P(s)$$

then

$$C(s)=P(s)N(s)-P(s)H(s)G_2(s)C^*(s)D^*(s)$$

Therefore

$$C^*(z)=PN^*(z)-PHG_2^*(z)C^*(z)D^*(z)$$

$$C^*(z)=\frac{PN^*(z)}{1+PHG_2^*(z)D^*(z)} \tag{69}$$

References

1. Sampled-Data Processing Techniques for Feedback Control Systems, Arthur R. Bergen, John R. Ragazzini. *AIEE Transactions*, vol. 73, pt. II, Nov. 1954, pp. 236–44. Discussion by John M. Salzer, pp. 244–45.

2. Factors in the Design of Digital Controllers for Sampled-Data Feedback Systems, J. E. Bertram. *Technical Report T-8/B*, Electronics Research Laboratories, Columbia University, New York, N. Y., Aug. 1955.

3. Analysis of Control Systems Involving Digital Computers, W. K. Linvill, J. M. Salzer. *Proceedings*, Institute of Radio Engineers, vol. 41, no. 7, July 1953, pp. 901–06.

4. The Analysis of Sampled-Data Control Systems, J. R. Ragazzini, L. A. Zadeh. *AIEE Transactions*, vol. 71, pt. II, 1952, pp. 225–34.

5. The Theory of Pulse Monitored Servomechanisms and Their Use for Prediction, R. H. Barker. *SRDE Report No. 1046*, Signals Research and Development Establishment, Christchurch, England, Nov. 1950.

6. Automatic Feedback Control System Synthesis (book), J. G. Truxal. McGraw-Hill Book Company, Inc., New York, N. Y., 1955.

7. Fundamental Theory of Servomechanisms (book), L. A. MacColl. D. Van Nostrand Co., Inc., New York, N. Y., 1945.

Discussion

E. I. Jury and W. Schroeder (University of California, Berkeley, Calif.): The authors are to be commended for their novel approach of compensating continuous feedback systems by introducing the digital compensator in the parallel path. It is noteworthy to mention, as indicated by the authors, that in such a configuration the continuous system will still operate (assuming it is stable) upon failure of the digital computor. The limitation of this configuration, in comparison to digital compensation in the series path, as recognized by the authors, is that the transient response depends on when the input signal is applied relative to the sampler.

The over-all objective of synthesis introduced by the authors is to obtain a desirable response of the continuous system at the sampling instants only, without due emphasis on the actual response between the sampling instants. The uncertainty of the use of the z-transform in some cases may yield a very undesirable continuous response, although the error may be zero at the sampling instants.[1]

In our opinion, the design procedures could be considerably improved by using the modified z-transform,[2,3] which yields the continuous response. Consider first the analysis procedure based on the modified z-transform, as follows: Equation 51 of the paper can be rewritten in the form

$$C(s)=A(s)H(s)D^*(s)[R^*(s)-C^*(s)]+A(s)R(s) \quad (70)$$

The modified z-transform of equation 70 yields

$$C^*(s,m)=AH^*(s,m)D^*(s)[R^*(s)-C^*(s)]+AR^*(s,m) \quad (71)$$

where $C^*(s)$ is given by equation 53 of the paper. The terms $AH^*(s,m)$ and $AR^*(s,m)$ can be obtained from existing tables of modified z-transforms.[2,3] The inverse formula which yields the actual response is presented by[3]

$$C(n,m)T=\frac{1}{2\pi i}\int_\Gamma C^*(s,m)z^{n-1}dz \quad (72)$$

Since m is considered constant in the integration, the response can be obtained by tables,[4] long division,[5,3] or by the sum of the residues. To obtain the continuous response, let m vary from zero to unity, and when m is unity the response at the sampling instants is obtained.

The modified z-transform method has been applied to example I introduced by the authors and the complete response, given by equations 71 and 72 is

$$C(n,m)T=\frac{1}{2\pi i}\int_\Gamma \times \left[\frac{0.764z^2-0.125z+0.361}{z^2(z-1)}-0.787e^{-0.575m}z^{-1}\cos(1.455m-14.5^\circ)+0.541e^{-0.575m}z^{-2}\cos(1.455-1.455m+14.5^\circ)\right]z^{n-1}dz \quad (73)$$

Equation 73 can easily be evaluated to yield

$$t=0,\ C=0 \quad (74)$$

$$0\le t\le T,\ C=0.764-0.787e^{-0.575m}\times\cos(1.455m-14.5^\circ) \quad (75)$$

$$T\le t\le 2T,\ C=0.639+0.541e^{-0.575m}\times\cos(1.455-1.455m+14.5^\circ) \quad (76)$$

$$2T\le t\le\infty,\ C=1.0 \quad (77)$$

This response has been plotted in Fig. 7, where it is noticed that it coincides exactly with the calculated response of the authors at the sampling instants. It is also observed that this response differs slightly from the experimental curve, Fig. 4(C). This discrepancy may be due to imperfect cancellation of poles in the experimental setup.

The sensitivity of the response to shifts in gain, sampling period, or dynamical characteristics causing imperfect cancellation, not discussed by the authors, is an important practical consideration. It should be indicated that the design of a practical system requires the selection of the proper compensating parameters that not only satisfy the desired specifications but also yield a response which is sufficiently insensitive to shifts.

Since the system to be compensated is continuous, the choice of the sampling period is important and should have been discussed in more detail by the authors. For instance, the choice of the sampling period of the digital compensator depends on the desired insensitivity of the response to shifts within the system, as well as the desired settling time and the practical limitations of realizing the sampler and the compensating device.

In addition to applying the modified z-transform method to analysis, it can also be applied to design procedures. The integrand in equation 72 can be easily forced to yield a response with zero error after a finite time, from which the compensator transfer function $D^*(z)$ is determined.[4] This procedure eliminates the uncertainty of the system behavior between the sampling instants and removes the necessity of choosing the "best" over-all transfer function. This procedure can also be advantageously applied to systems with arbitrary delays or highly oscillatory systems, in which case the z-transform method may yield misleading results.

In conclusion the authors have introduced a valuable approach for compensating continuous systems and their method of design, in our opinion, can be improved if consolidated with the modified z-transform approach.

References

1. Hidden Oscillations in Sampled-Data Control Systems, E. I. Jury. *Report 155*, Electronics Research Laboratory, University of California, Berkeley, Calif., series 60, 1956.

2. Synthesis and Critical Study of Sampled-Data Control Systems, E. I. Jury. *AIEE Transactions*, vol. 75, pt. II, 1956 (*Paper No. 56-20f*).

3. The Pulse Transfer Function and Its Application to Sampling Servo-Systems, R. H. Barker. *Proceedings*, Institution of Electrical Engineers, London, England, vol. 99, pt. 4, 1952.

4. Analysis and Synthesis of Sampled-Data Control Systems, E. I. Jury. *AIEE Transactions*, vol. 73, pt. I, 1954, pp. 332–47.

5. Additions to Z-Transform Theory, G. V. Lago. *Ibid.*, pt. II, pp. 403–08.

6. Discrete Compensation of Sampled-Data and Continuous Control Systems, E. I. Jury, W. Schroeder. *Report 154*, Electronics Research Laboratory, University of California, Berkeley, Calif., series 60, 1955.

K. K. Maitra and P. E. Sarachik: We wish to thank Dr. Jury and Dr. Schroeder for their interesting and constructive comments.

The discussers raise the possibility of obtaining a very undesirable envelope function for the continuous output, if the synthesis is done solely based on the information available at the sampling instants. This difficulty is not, however, an unique one in the present problem. It is inherent in the synthesis of all sampling servos, and the system considered in the present discussion is only a special case. Methods[1,2] similar to the modified z-transform, suggested by the discussers, were available to the authors at the time this investigation was carried out. It is recognized[3] that a more sophisticated synthesis procedure would be desirable which would optimize the continuous response rather than the response at the sampling instants. The method suggested by the discussers would obviously lead to a more complicated structure for the digital compensator. It has been shown by the authors in the section entitled "Alternative Method of Design" that it could be possible to obtain a desirable response by simply adjusting the coefficients of the transfer function of $D^*(s)$. This method would be easier to carry out if the system is studied on the computor, which would undoubtedly be the ultimate test, before any system is put into operation.

It should be understood that the method of synthesis suggested in the paper shall serve as a guidepost to the design of an actual system. In our opinion, the modified z-transform method could be more conveniently used to compute and check the response of the system during the inter-sampling periods, rather than used for the purpose of the preliminary design. This may be justified in view of the relative simplicity of the recommended method.

The discussers suggest that "the integrand in equation 72 can be easily forced to yield a response with zero error after, a finite time, from which the compensator transfer function $D^*(z)$ is determined." What they suggest leads to choosing a $KR^*(s)$ quite independently. It is not clear to us, how an independent choice of the response function could be possible, when there are undesirable zeros present in the pulsed transfer function $AH^*(z)$. In fact one of the main objects of this paper

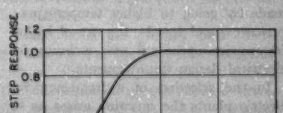


Fig. 7. Continuous response of example I

is to show that such choice can not be made independently, but is subject to certain limitations.

The response, plotted in Fig. 7, differs slightly from the authors experimental curve, Fig. 4(C). This discrepancy is due to the fact that there is about 5 per cent delay present in the digital controller used for this study. This factor has been ignored in the computation.

We consider that the selection of the sampling period T should be done depending on the performance of the uncompensated system. As such, this is discussed in the illustrative example I as given in the paper.

In conclusion, a complete synthesis procedure based on the modified s-transform method, as suggested by the discussors would be highly useful.

References

1. Automatic and Manual Control (book). "Characteristics of Sampling Servo Systems," C. Holt Smith, D. F. Lawden, A. E. Bailey, and discussion by R. H. Barker. Butterworths Scientific Publications, London, England, 1952.

2. Analysis of Errors in Sampled-Data Feedback Systems, J. Sklansky, J. R. Ragazzini. *AIEE Transactions*, vol. 74, pt. I, May 1955, pp. 65–71.

3. See reference 1 of the paper.

Shipboard Use of 400-Cycle Electric Power

J. M. APPLE
ASSOCIATE MEMBER AIEE

E. W. LUSBY
NONMEMBER AIEE

IN any discussion of 400-cycle electric power systems, the question comes to mind of why frequencies should be considered other than the established standard 60-cycle frequency now widely and most commonly used in electric systems ashore and afloat. Since 1932, 60-cycle a-c systems have been used in the majority of the ships' service electrical installations, particularly installations of large capacity. These have operated satisfactorily, and they served well during World War II. However, during and since this period the trend has been to increase the use of electric power, thereby increasing weight and space requirements of equipment out of proportion to the size of the ship, and to require improved performance of the equipment installed. The answer to this question can be summarized by the word "progress." The purpose is to obtain better performance of electric equipment, or the same performance with less material, weight, and space. To achieve these goals, consideration must be given to the basic factors which influence electric equipment designs, one of which is frequency. Just as in other engineering fields where progress is being made by going to higher temperatures, pressures, and speeds, the use of higher frequencies opens up new avenues for electrical design improvements.

In the selection of a frequency for electric plants the question arises as to the use of 400 cycles rather than some intermediate or higher frequency, particularly some multiple of 60 cycles. In the range of 400 cycles it has been determined that:

1. Generator synchronous speeds closely match available turbine prime mover speeds.
2. Efficient motor designs are possible with speeds matching present maximum speeds of driven auxiliaries.
3. Cable voltage drops are not excessive nor is cable current-carrying capacity appreciably limited.
4. Efficient design of magnetic devices such as transformers, magnetic amplifiers, and synchros is possible.

The selection between 400 cycles and a multiple of 60 cycles (420 or 360) is not so definite. The use of 400 in lieu of 420 cycles does permit the application of 60/400-cycle motor generator sets with induction motor drive which can be controlled closer to a 400-cycle output than would be possible with a rating of 420 cycles. Since considerable progress has been made in the aircraft industry on 400-cycle systems and equipment, this frequency is considered a logical choice from a standardization viewpoint. As a result of extensive military application of 400-cycle power the U.S. Defense Department has included 400 cycles in Military Standard 205.

Discussion

The use of higher frequency electric plants such as 400-cycle plants, offers certain advantages which are very attractive, particularly in installations where weight and space requirements are of importance. An increase in system frequency increases maximum speed of rotation, thereby decreasing the weight of rotating electric machinery. In addition, high-frequency transformers and reactors require a lower value of maximum flux, and hence a smaller iron core than corresponding 60-cycle units. Inherently smaller time constants permit increased speed of response, which is an advantage in certain control systems.

As is the case in most efforts to advance, there are other factors to consider which partially offset the advantages. Inherent characteristics of small, lightweight high-frequency equipment require that special consideration be given in the equipment design to noise reduction, higher operating temperatures, and mechanical problems that arise with the higher rotational speeds and smaller air gaps. The smaller heat dissipation surface requires that more consideration be given to ventilation and equipment protective features to prevent burn-out. Also, the advantages offered by the use of a higher frequency in marine applications have to be great enough to compensate for the high costs associated with the development, design, and manufacture of electric equipment for a 400-cycle frequency.

The foregoing paragraphs briefly outline some of the factors involved in the comparison of standard 60-cycle electric equipment with higher frequency equipment. Now, the effects of system frequency on specific electric apparatus will be considered.

Rotating Electric Machinery

An increase in system frequency increases maximum speeds of rotation, thereby decreasing weight of rotating electric machinery as compared to 60-cycle machinery. Figs. 1 and 2 illustrate the effects of increased operating speed on induction motors as compared to 60-cycle motors of the same rating.

Paper 56-184, recommended by the AIEE Marine Transportation Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Winter General Meeting, New York, N. Y., January 30–February 3, 1956. Manuscript submitted October 27, 1955; made available for printing November 30, 1955.

J. M. Apple and E. W. Lusby are with the U. S. Navy Department, Washington, D. C.

All opinions expressed in this article are the personal engineering opinions of the authors and do not necessarily represent the official views of the U. S. Navy Department.

Fig. 1. Comparison of a 30-hp 60-cps 1,200-rpm motor with a 30-hp 400-cps 12,000-rpm motor

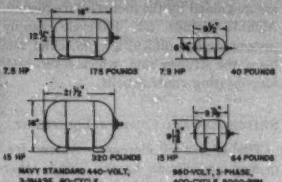


Fig. 2. Comparison of standard 60-cps motor with 400-cps motors

The primary objective in the development of 400-cycle motors has been to design for a synchronous speed of 6,000 to 8,000 rpm in those applications where the driven auxiliary may be driven at these speeds. These applications include ventilation fans and certain impeller-type pumps. Where lower auxiliary speeds are necessary, such as in compressor applications, 12,000-rpm motors and reduction gears are used.

The efficiencies of 400-cycle motors as developed for marine service are about the same as conventional 60-cycle induction motors. Although for a given magnetic material the core loss is proportional to a power of the frequency, the core losses are about the same as that for 60-cycle motors of the same horsepower rating. This is true because the best grade of silicon steel available was used and the thickness of laminations was reduced.

The 400-cycle motor will run hotter because of the smaller surface to dissipate heat. Thus, the 400-cycle motor requires more ventilating air properly channeled and directed. Also, since rotors are more susceptible to burn-out during momentary stalls as the frequency increases, the rotor must be properly constructed and criteria for safe rotors had to be established. Calculations made by manufacturers indicated that if all heat was stored in the rotor bars, the rotor would be destroyed in 12 seconds with six times full-load current. This procedure has been used for the selection of overload relays. On actual test, the motor withstood the stalled rotor test 30 seconds with nine times full-load current.

The noise level of 400-cycle motors is of course higher than standard 60-cycle motors because of the higher speeds and the need for more cooling air. The noise level can be controlled through the proper use of sound-absorbing material and proper installation.

Table III. Comparison of 60-Cycle 80 C and 400-Cycle 130 C Rise Transformers

Kva	Weight, Pounds		Volume, Cubic Inches	
	60 Cps	400 Cps	60 Cps	400 Cps
5	90	30	1,450	600
15	205	75	4,000	1,750
25	280	105	5,700	2,150
50	510	165	10,025	3,600
100	950	310	17,300	6,060

Steam and Gas Turbine Generators

The effects of an increase in system frequency on electric generators is of prime importance, and therefore selection of the type of generators was one of the first items to be considered. Because of the increase in maximum available speed of rotation, reduction gears can be eliminated from steam and gas turbine drive generators. This, of course, resulted in a further decrease in weight of the combined unit. To illustrate the effect of increased system frequency on steam turbine generator size and weight, refer to Table I, which gives comparison data on a 400-kw 60-cycle ship service steam turbine generator used in past destroyers and a 600-kw 400-cycle ship service generator used in the *USS Timmerman*. As can be readily seen, the 400-cycle generator is approximately one-third the weight of a 60-cycle unit and has a 50-per-cent (%) increase in capacity.

Since both gas turbine and diesel generators are suitable for emergency service, efforts have been directed toward the development of gas turbine drives for emergency service. Also, development of lightweight diesel engines was undertaken as a parallel effort and with some success. Table II gives comparison data on emergency diesel generators used on past destroyers and gas turbine and lightweight diesel generators used on the *USS Timmerman*.

Table I. Comparison Data on Steam Turbine Generators

Generator Characteristics	Past Destroyers	USS Timmerman	
		Generator No. 1	Generator No. 2
Generator type	synchronous	synchronous	synchronous
Generator rating	400-kw 80% pf*	600-kw 80% pf	600-kw 80% pf
	450-volt 60-cycle	1,000-volt 400-cps	1,000-volt 400-cps
	3-phase 1,200-rpm	3-phase 12,000-rpm	3-phase 24,000 rpm
Turbine type and rating	6-stage 525-psig†	7-stage 815-psig	6-stage 1,805-psig
	825 F‡ TT§	1,045 F TT	1,040 F TT
	10,059-rpm	12,000-rpm	24,000 rpm
Weight, pounds			
Generator	4,350	1,640	1,800
Turbine	6,375	1,950	2,000
Auxiliaries	11,990	3,110	2,753
Total, dry	22,615	6,700	6,553
Type of excitation	50 kw d-c, oversize to supply d-c loads	single stage rotary amplifier	static

* Power factor.
† Psi gauge.
‡ Fahrenheit.
§ Throttle temperature

Table II. Comparison Data on Diesel and Gas Turbine Generators

Generator Characteristics	Past Destroyers	USS Timmerman	
		Emergency Generator No. 1	Emergency Generator No. 2
Prime mover type	8-cylinder 2-cycle Diesel	gas turbine	4-cylinder radial 2-cycle Diesel
Prime mover rpm	1,200	20,000	1,800
Generator rating	100-kw 80% pf	250-kw 80% pf	250-kw 80% pf
	450-volt 60-cycle	1,000-volt 400-cps	1,000-volt 400-cycle
	3-phase 1,200-rpm	3-phase 8,000-rpm	3-phase 12,000-rpm
Total weight, dry	9,300 pounds	1,585 pounds	8,500 pounds

Conversion Equipment

Increasing system frequency has a profound effect on transformers and reactors since the lower value of maximum flux permits a smaller iron core. As shown in Table III, weight reductions in the order of 60% may be expected. The main problem connected with 400-cycle transformers was the discomforting hum created by the magnetic circuit. Loudness of sound as detected by the hum ear is a function of intensity and frequency. Sound intensity and loudness differ in that intensity is a measurement of energy at a certain distance from the source, while loudness is the sensation to the brain caused by the sound source. The 800- to 1,600-cycle vibration encountered at 400-cycle operation is within a frequency band to which the ear is particularly sensitive. Therefore particular efforts were made to reduce transformer noise to a minimum. This problem is being overcome by changes in the type of core construction.

To supply power to the d-c loads such as degaussing and battery-charging, numerous methods are available which include direct-driven generators, motor generators, mercury arc rectifiers, and copper oxide and selenium rectifiers. Direct-driven generators are not used because of the commutation problems associated with the higher speeds. Motor generators with geared generators do not offer sufficient advantage over static rectifiers to compensate for the increase in weight. Selenium rectifiers are used on the basis of weight savings, lower maintenance, and higher efficiency.

Interior Communication and Electronic Equipment

An increase in system frequency also has considerable effect on interior communication and electronic equipment. As with other equipment, the reduction in size and weight is manifested mainly in the reduction in size of chokes, filters, power, and filament transformers. The use of 400-cycle power offers both advantages and disadvantages depending on the interior communication system being considered. In the case of the salinity indicator system the use of a 400-cycle supply is superior to a 60-cycle source, since polarization effects are reduced and accuracy is increased. Gyro compass systems also commonly use 400-cycle power. One of the difficulties tending to retard widespread use of 400-cycle synchros is the necessity for carefully balancing reactances in wiring between secondaries. The errors due to unbalanced reactances in cable runs are magnified by the increased frequency. Audible signals for 400-cycle supply present design difficulties, the chief one being the striking rate, which is too high for vibrating contact bells.

Switchgear and Control Equipment

The effects of increased system frequency on circuit breakers and other switchgear components were determined from tests conducted at the manufacturer's plants and at Naval laboratories. Based on these tests, standard 500 volt 60-cycle frame sizes and current ratings are satisfactory for 400-cycle 500-volt application. Also, interrupting tests were performed but only on a limited scale due to the capacity of 400-cycle test stations. Because of this limitation, standard 60-cycle molded case breakers, which have a minimum 60-cycle interrupting capacity of 15,000 amperes, are only approved for applications where available fault currents are 10,000 amperes or less. However, the tests have indicated that the circuit breakers would most likely possess the same interrupting capability at 400 as at 60 cycles per second (cps). Since the increase in frequency does not greatly influence the size and weight of other switchboard components, such as instruments and switches, the over-all switchboard dimensions and weight are not materially different from those of 60-cycle switchboards.

The influence of 400-cycle frequency on motor control equipment required new design overload relays and contactor-operating magnets. Since the majority of the weight and space savings result from the higher operating speeds of rotating machinery, control efforts had been concentrated on obtaining motor controls that operate satisfactorily from a 400-cycle supply while maintaining size and weight approximately the same as for 60-cycle control equipment.

Numerous protection problems arise in the consideration of 400-cycle supply frequency. Perhaps the most important one concerns the motor overload elements. Since the starting current of 400-cycle motors is higher than that of 60-cycle motors, and since the 400-cycle motors are smaller with less area to dissipate heat, the motor overload elements are required to operate faster. Therefore, overload elements had to be developed for 400-cycle application. Also, overcurrent devices for circuit breakers required calibration for the 400-cycle frequency. The determination of the circuit-breaker trip settings was also difficult because of the limited data available on circuit-breaker operation at 400 cycles and 400-cycle machine characteristics.

Cable

Power and lighting cable is affected by an increase in frequency in two ways: first, the current rating is reduced from 10 to 30% for cables, 200,000 circular mils and over; and, second, the voltage drops are increased because of the increased cable impedance, particularly in the larger cables. Therefore, 400-cycle system cable weights are more than 60-cycle weights at the same voltage.

Lighting Equipment

The use of fluorescent lamps with a 400-cycle supply also results in weight and space savings. This reduction is attributable to the smaller 400-cycle ballast which is approximately one-fourth the size of a 60-cycle ballast and weighs 50% less. The smaller sized ballast also gives more freedom to the fixture designer, in that shallower fixtures can be built for installations in areas having low overhead. In a typical ship installation lighting fixture weights can be reduced approximately 10%. However, if frequency conversion equipment is necessary the weight and space savings may be offset by the additional equipment. The use of 400-cycle power supply also improves fluorescent lamp efficiency and reduces stroboscopic effect.

Past and Present Shipboard Applications

In 1947, with the development of 400-cycle power applications in aircraft, fire control equipment, and radar, consideration was first given to the use of 400-cycle power aboard ship. At the start, conversion equipment was incorporated as part of the basic components for fire control and electronic applications. In 1948, the first installations of central power conversion plants were provided on shipboard for aircraft servicing and testing facilities. At about this time, 400-cycle power was adopted for gyro drives in lieu of the special frequencies previously used.

Present applications using 400-cycle power on U.S. Navy ships include gyros, certain fire control and electronic applications, aircraft testing and servicing and some special control circuits. There are no general power applications being made at present, although in some cases where a 400-cycle power source is pro-

Table IV. Preliminary Weight Summary of Electric Plant, USS Timmerman

Component	Weight in Thousands of Pounds				
	60-Cps 450-Volt Conventional Plant	60-Cps 450-Volt Improved Plant	400-Cps 450-Volt Plant	400-Cps 600-Volt Plant	400 Cps 1,000-Volt Plant
Ship service generators, exclusive of prime movers, 600 kw each	20.0	12.3	5.4*	11.1†	5.4‡
Emergency generators, 250 kw each	8.9	6.2	2.5	4.5	2.1
Switchgear	19.4	17.1	22.6	17.1	16.5
Panels	6.6	7.4	7.1	7.1	7.1
Motors and controls	52.2	25.4	11.7	14.7	16.1§
Transformers	3.1	2.6	1.9	1.2	0.9
Cable	72.2	72.2	77.2	68.1	59.5
Tubes and wireways	20.7	20.7	22.1	20.5	17.4
Supplementary transformers					0.4‖
Total	203.1	162.9	150.5	144.3	127.0
Percentage savings over conventional 450-volt 60-cycle plant		19.8	25.9	29.0	37.5

* 24,000 rpm synchronous generator design.
† 12,000 rpm synchronous generator design.
‡ 24,000 rpm induction generator design.
§ Motor control weights include capacitors for pf correction.
‖ Transformers for heaters and motors below 2 hp.

vided for the applications previously mentioned it may be used for special installations. For example, on aircraft carriers where a considerable amount of 400-cycle power is required for aircraft servicing, this same power supply may be used for aircraft starting by converting it to direct current. The advantage in this case is the reduction in size of the necessary conversion equipment using 400-cycle power in lieu of 60-cycle power. Another example is the use of already available 400-cycle power to supply the arc discharge type of searchlights since the auxiliaries, such as the ballast, can be made smaller.

Since 400-cycle power is not being generally used for shipboard applications at present, each application is treated on a case basis. In other words, both the advantages to be gained in the design and performance of the power-utilizing equipment and any disadvantages from the system and power generation or conversion equipment standpoint are evaluated, and the selection of 400 or 60 cycles is made on the basis of the over-all evaluation.

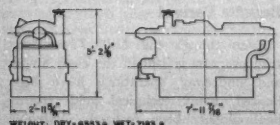


Fig. 3. Outline of a 600-kw 1,000-volt 400-cps 24,000-rpm synchronous generator driven by a 2,000-psi, 1,050 F steam turbine

Experimental Ship USS Timmerman

Realizing the advances that could be achieved by a complete investigation of such items as increased electric plant frequency, the U.S. Navy decided to build a destroyer of advanced design in which the maximum number of new design components and systems would be incorporated. This ship is the *USS Timmerman*, formerly the *DD828*, an now the *EAG152*. World War II destroyer experience indicated the need for a ship having a greater cruising radius, increased speed, and improved seakeeping characteristics. Therefore, the main feature of the experimental destroyer was a 100,000-shaft-horsepower (SHP) machinery plant of equal or less weight and space than the 60,000-SHP plant of the *DD692*-class World War II destroyer.

In attempting to achieve this goal the following items were considered worthy of consideration:

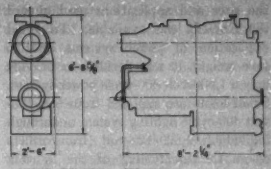


Fig. 4. Outline of a 600-kw 1,000-volt 400-cps 12,000-rpm synchronous generator driven by a 815-psi, 1,045 F steam turbine

1. Reduction of design margins.
2. Elimination of pyramiding of design margins.
3. Reduction in safety factors. Each component design was to be analyzed to make certain that safety factors were consistent.
4. Use of higher rotative speeds.
5. Use of higher voltage and frequency.
6. Use of better materials, taking into consideration critical materials which present a bottleneck during emergencies.

With the foregoing guidelines established, there followed a period during which numerous design studies were made in order to form a basis for the selection of the most promising electric plant characteristics. Leading manufacturers of generators, switchgear, transformers, motors, control equipment, and auxiliaries were called upon to design, develop, and furnish equipment which would meet the requirements of the basic concepts. As a result, a 1,000-volt 400-cycle electric plant was selected for the *USS Timmerman* on the basis of estimated data furnished the design agent. The estimated weight and space data for the principal components are given in Table IV. The estimated data were submitted largely on the basis of original new design, since comparatively little manufacturing experience on higher voltage and frequency equipment, of

Table V. Final Weight and Space Summary of Electric Plant, USS Timmerman

Component	Weight, Thousands of Pounds		Space, Cubic Feet	
	450-Volt, 60-Cps	1,000-Volt, 400-Cps	450-Volt, 60-Cps	1,000-Volt, 400-Cps
Generators, exclusive of prime movers; two 600-kw ship service and two 250-kw emergency	29.3	6.1	100	40
Switchgear	19.4	21.0	520	595
Panels and bus transfers	7.5	9.0	114	135
Motors and controls	47.2	23.2	570	283
Transformers	9.1	2.7	90	30
Cable	69.8	59.5		
Tubes and wireways	20.3	17.4		
Total	202.7	138.9	1,304	1,085
Percentage savings over 60-cycle 450-volt plant		31.5		22

the sizes and applications used aboard the *EAG152* was available. To avoid duplication of power systems and to save weight it was decided to provide only a 1,000-volt 400-cycle power system. Small loads are supplied from the 120-volt 400-cycle lighting system or through individual 1,000/120-volt transformers where the functioning of the lighting system could be impaired. Certain loads, such as heaters, galley equipment, etc., were supplied from individual transformers at their required voltages because of the excessive development costs involved for obtaining equipment suitable for operation at 1,000 volts. The decision to use 1,000 volts and 400 cycles in the *EAG152* took into account the fact that the design is not a prototype but an experimental advanced design offering an opportunity to explore the advantages of higher voltage and higher frequency electric plants. One of the disadvantages of a 1,000-volt 400-cycle electric plant is that shore power facilities at this frequency and voltage are not available at the various shore installations. Accordingly, the *USS Timmerman* was provided with a shore power motor-generator set. The generator rating is 250 kw, 1,000 volts, 400 cycles. The motor rating is 440 volts, 60 cycles, 3-phase.

The electric generating plant on the *USS Timmerman* consists of two 600-kw synchronous turbogenerators. One unit, illustrated in Fig. 3, is driven by a 2,000-pounds per square inch (psi), 1,050-degree-Fahrenheit steam turbine and is a "package unit," direct-coupled, turning at 24,000 rpm, and having a static excitation system. The other unit, illustrated in Fig. 4, is driven by an 815-psi, 1,045-degree-Fahrenheit turbine of the package-unit design, direct-coupled, turning at 12,000 rpm and having a rotary exciter. In addition to these there are two 1,000-volt 400-cycle 3-phase emergency generators. The forward unit is driven by a gas turbine while the after unit is driven by a diesel engine. All pumps and other auxiliary machinery, with the exception of turbine-driven feed pump and main forced draft blowers, are driven by 950-volt 400-cycle motors. These high-speed motors utilized silicone insulation, which is inherently resistant to moisture and can be operated at higher temperatures.

Representative figures as to actual space and weight savings in the complete electric power and lighting system resulting from using 1,000 volts and 400 cycles in lieu of 450 volts and 60 cycles on the *USS Timmerman* are given in Table V. Approximately 15% of the total weight savings is attributable to modern design, 5% to the increase from 450 volts to 1,000 volts, and the remaining 11.5% results from the increase in frequency.

Manufacturing Problems

The selection of a 1,000-volt, 400-cycle plant for the *USS Timmerman*, while offering an opportunity to explore the advantages of higher voltage and frequency, also presented many engineering and manufacturing problems. The lack of experience with 400-cycle electric equipment in the ratings required necessitated development of practically all components from the generator to the driven auxiliary. For example, test motors had to be built to enable the manufacturer to determine design guides for 400-cycle motors. The higher rotative speeds and higher temperatures presented a difficult bearing problem and other mechanical problems which still have not been solved. In addition, electrical characteristics such as rotor design, motor insulation system, and safe rotor conditions had to be determined. At the same time similar development work was taking place on motor controls, generators, transformers, and circuit breakers. The over-all development was of large proportions and from them came the *Timmerman* 400-cycle electric plant.

Operational Problems

As was to be expected, during the 3-year period since completion of the *USS Timmerman* numerous operational difficulties have occurred. Some of the more significant failures are as follows.

Generators

Operating experience with the rotary amplifier exciter on the no. 1 12,000-rpm generator indicated that the shaft would score after approximately 800 hours of operation. The rotary amplifier exciter and regulator has since been replaced with a static exciter and regulator to overcome this problem. Extensive modifications to the no. 2 24,000-rpm generator were required soon after completion of the ship. The capacity of the generator was found to be less than rated because of short circuited field coils and arcing at the slip rings. Modifications included rewinding the rotor and increasing field turns, and installing new slip rings and brush holder assembly which accommodated additional brushes, and reduced current density. This generator performs satisfactorily except for limited brush life.

Operating deficiencies have occurred with the emergency generators. The high ambient temperature in the emergency gas turbine-generator compartment reduced the output of the machine by approximately 30%. Alterations to the ventilation system are being made to correct this deficiency. The diesel emergency generator has had a history of bearing failures on the upper end bearing of the generator, which was formerly a grease-lubricated bearing. This bearing has been replaced with an oil-lubricated bearing, which should eliminate this type of casualty.

Motors

The 950-volt 400-cycle motors are smaller, have less thermal capacity and they operate at higher speeds and temperatures than are normally encountered. It is not surprising, therefore, that numerous failures have occurred. It is believed that a solution to these and other motor problems will be achieved as a result of the following program now in progress:

1. Use of heat-stabilized silicone-lubricated antifriction bearings which are good for temperatures up to 150 degrees centigrade (C).
2. Replacing, in some instances, relatively flimsy end bells of stamped construction with more rigid end bells of cast construction.
3. Paying particular attention to bearing fits and tolerances, rotor balancing, alignment of motors, and end bell fits, all of which are critical. Increasing difficulty has been experienced with balancing of the high-speed rotating equipment which is at least partly responsible for early bearing failures.

Controls

There have been a considerable number of motor controller casualties because of failure of relays, control transformers, and contactors. Remedies that are being effected to correct these failures are:

1. Modification of controller arc chutes, and contacts.
2. Modification of the thermal overload relays to insure more positive overload protection.

Benefits Derived

Even though the *USS Timmerman* trial program is still in the early stages and underway operations have been limited, many technical achievements have already been realized. Valuable information from shipboard operating experience and tests has already been incorporated into the design of 400-cycle turbogenerators for limited applications

in new ships under construction. The new carriers, for example, have a 1,000-volt 400-cycle system for aircraft starting and servicing.

Additional operating experience on the electric equipment aboard the *USS Timmerman* is still needed in order to obtain information as to life, amount of maintenance required, dielectric properties, and conditions of bearings and insulation resulting from continuous service.

Summary

Over the past years shipboard electric installations have grown from simple low-voltage (120 to 230) d-c systems to large complex systems. This growth has advanced in most cases, from the d-c systems to 450-volt 60-cycle a-c systems. There is no reason to expect that the shipboard electrical installations will not continue to grow, with further demands for still better performance and more output with less space and weight. A continuing increase may therefore be expected in the use of 400-cycle power. If this occurs, it is expected that it will not be accomplished by a complete change in the basic electric plant at one time as was the case in the change from direct to alternating current, but by a gradual change with an increasing portion of the plant made up of 400-cycle applications.

This increase will probably be more rapid in the field of electronics and control circuits than in other fields. However, 400-cycle fluorescent lighting installation will probably be used in the not too distant future. In the power drive field, it is expected that increasing use will be made of 400 cycles where increased speeds are desired to obtain better characteristics on the part of driven auxiliaries.

This growth in 400-cycle power utilization will necessitate larger 400-cycle power generation and conversion plants, together with the necessary distribution systems. The general application of 400-cycle power should result in the development of new and improved bearings and gears, new types of overload protection, and changes in magnetic materials and the design of magnetic structures. No basic changes in distribution equipment such as cable, fittings, and switchgear will be required.

Discussion

J. D. Shuster (Bethlehem Steel Company, Quincy, Mass.): The turbogenerators in this experimental ship's installation were both noteworthy in being direct-drive rather than geared, one running at 24,000 rpm (2-pole), the other 12,000 rpm (4-pole). It is interesting to note that the 12,000-rpm unit originally used a rotating exciter but has since been provided with a rectifier for excitation, as originally provided for the 24,000-rpm unit. Other than for weight comparisons, nothing is said about the induction generator turbine-driven set built in conjunction with the *Timmerman*. It would be profitable to have a report on any tests made on it.

We should like to know the per cent of auxiliaries other than fans which were direct-motor drive as against those which were geared. The saving in auxiliary weight is not tabulated but could be credited (or charged) to the use of 400-cycle power.

Table V and statements in the section describing the ship indicate that 400-cycle equipment allows a saving in weight of 11½% over standard Navy equipment and that other new features (probably such as class-*H* insulation) allow a saving of 15% over standard Navy equipment. These percentages cover the whole electrical installation and it appears that their relative value would be different when applied only to the motors. The comparison of the two 30-hp 12,000 rpm motors pictured in Fig. 1 showing an 88.5% weight saving is certainly startling, even though the class of insulation is different, particularly when it is realized that the small one is gear-in-head.

The motor and control item for 400 cycles, 1,000 volts, in Table IV mentions capacitors for power factor (pf) correction. We see no other mention of pf correction and assume that the actual *Timmerman* installation did not use capacitors for this purpose.

The statement concerning calculated and actual induction motor squirrel-cage rotor abilities relative to heavy overload currents is interesting in relation to the assessment of the use of 60-cycle multispeed induction motors for deck machinery. If the squirrel-cage motor rejection for such use is equally conservative, the industry may be missing a satisfactory solution to the a-c deck machinery problem. It is a fact that 11 ships have been ordered from European yards using dual multispeed induction motors for cargo winches and that two of them owned by the Hamburg-American Line are now calling at our Gulf Coast ports.

The discussion of motor controls emphasizes that starting currents are higher and heat dissipation is more difficult in the 400-cycle motors, resulting in the need for faster operating overload relays. It would appear that this would result in many nuisance trips, particularly on starting.

We assume that the transformer weights in Table III for 60 cycles are for 450-volt primary, and for 400 cycles 1,000-volt primary, and from Table IV, columns 3 and 4, the estimated weight saving, due solely to the differences in frequency, is 26.8%.

The information concerning fluorescent lights is noteworthy and, as the paper implies, might be one of the early applications for 400-cycle power for merchant vessels. It is understood that at least one land application has been made, in a gymnasium.

We believe that the AIEE owes the authors a vote of thanks for their excellent paper covering a broad new type of marine application of electric power.

Arthur R. Gatewood (American Bureau of Shipping, New York, N. Y.): The following questions occur to us:

1. How practical is it to use present-day selenium rectifiers?
2. Would across-the-line starting be practicable and, if so, has it been given consideration?
3. Will it not be necessary to use special cables because of the voltage increase? The commonly used merchant standard has been developed for a maximum of only 600 volts.
4. While not mentioned, it would appear that the copper current density of the electric cables is the same as for a 60-cycle plant, and it would be of interest to know whether the skin effect of the higher frequency was considered.
5. With the use of increased speeds on the order of 6,000 to 8,000 rpm and up to 12,000 rpm, as mentioned, it is understood that more attention will have to be paid to bearing fits, alignments, and rotor balancing. Since turbochargers used in many present-day commercial applications of diesel engines are designed for speeds on the order of 40,000 to 50,000 rpm, it is wondered whether the authors intended to imply that there might be unusual problems in this connection.

V. W. Mayer (Maritime Administration, Washington, D. C.): The investigation of the use of the 400-cycle electric power system as described deals primarily with its use on naval ships where 400-cycle equipment has already found many applications. In the use of 400-cycle 1,000-volt equipment it has been established that a considerable saving in material, weight and space may be obtained. This leads to a question as to what would be the advantage in the use of 400-cycle electric power systems on ships of our merchant marine.

It would appear that a saving in material used in the construction of electrical equipment should reflect to some extent a reduction in the over-all cost of a ship, and this would certainly be attractive to any engineer. However, because of the apparent high design and development costs of 400-cycle equipment at this time there is some question as to whether an actual saving would result. In fact, it is not unreasonable to assume that the costs would be far in excess of the 450-volt 60-cycle conventional equipment. While the advantages of high-speed electric rotating equipment are recognized, it is doubtful if any advantages from the general use of 400-cycle equipment on merchant ships at present would be

sufficient to justify its high cost. It is also apparent that present maintenance costs on this equipment are extremely high and could not be considered acceptable.

There is no question as to the advantages in any saving which can be gained in a reduction in the use of critical materials during emergencies. The saving of any weight of electrical equipment on merchant ships is also important, since this could permit an increase in a ship's available deadweight capacity.

A saving in space on merchant ships which can be made available to pay load such as additional space for either cargo or passengers is always advantageous. Since the major electric equipment involved in space saving would be installed either in the engine room, fan rooms, winch control rooms, or electrical distribution centers, any space gained would naturally have to be obtained from these spaces.

It has been pointed out that dimensions and weights of switchboards and control equipment for 400-cycle equipment are approximately the same as for 60-cycle equipment. This would not permit any appreciable saving in space requirements of electrical distribution centers, with the exception of spaces where transformers are installed, where some space saving might be possible. While the size of electric motors can be reduced, there are many applications where the use of the high-speed 400-cycle motors cannot be applied directly to the driven auxiliary and where it becomes necessary to provide reduction gears. The combination of motor, reduction gears, and the auxiliary would therefore probably require nearly as much space as the conventional 440-volt 60-cycle motor.

Inasmuch as the switchboards and certain motors, in combination with reduction gears and the driven auxiliaries for 400-cycle equipment, will require approximately the same space as corresponding 60-cycle equipment, it would appear that, even though some space could be gained by the use of 400-cycle generators, transformers, and certain high-speed direct-connected auxiliaries, the over-all size of the engine rooms could not be materially reduced. In connection with direct-connected auxiliaries on the experimental installation, it would be of interest to know the specific applications where it has been possible successfully to utilize these motors without the use of reduction gears. It would also be of interest to know to what extent synchronous speeds of 6,000 to 8,000 rpm have influenced the characteristics of either the driven auxiliaries or the associated systems. It is not believed that the size of either wiuch control rooms or fan rooms can be reduced very much because of the space requirements for the control equipment and the lighting and power distribution panels which are occasionally installed in these spaces and which are approximately the same size as the 60-cycle equipment. It is possible that some space could be saved if the size of ventilation ducts and fans are reduced by the use of high-pressure ventilation. However, it is not current practice for marine installations in this country to provide such high-pressure systems; therefore, even though the size of the motors and fans could be reduced by the use of 400-cycle power, the size of the ventilation ducts would remain unchanged.

The use of 400-cycle power distribution on merchant ships up to the present has had little, if any, application and while the demand for electric power has consistently increased over the years, probably the weight and space requirements of the electrical equipment are out of proportion to the size of merchant ships.

Aside from the matter of saving material, weight, space, and cost, there is also the question of the reliability of this equipment in operation. The excessive failures of bearings, relays, control transformers, contactors, brushes, etc., such as have occurred on the experimental installation, would indicate that a highly trained crew of electricians, superior to that ordinarily available for merchant ships, would be required to maintain this equipment in proper operating condition.

In addition to the foregoing, because of commutation problems, etc., the use of high-speed motor-generator sets associated with modern cargo winch control would no doubt require considerable maintenance and thus cause delays in the handling of cargo. Such delays would be very costly to the ship operators, especially those on tight work schedules.

In so far as availability of shore power is concerned, the merchant ship is faced with the same problem as the navy ship, since 400-cycle shore power is not available.

In conclusion it may be observed that, of the total 31.5% saving in weight indicates, 5% is attributed to higher voltage, 15% to modern design, and only 11.5% to higher frequency. Thus, approximately 65% of the total saving in weight may be obtained without changing the frequency.

The purpose of this discussion is to bring out facts which must be considered in connection with the use of 400-cycle power on merchant ships as opposed to its use on Navy ships. It is not intended to discourage further study in the development of 400-cycle shipboard power. There is no doubt but that new and improved components, some of which can be made available for electrical systems operating on frequencies of other than 400 cycles will be developed as a result of the investigations described in this excellent paper.

J. G. Partlow (Westinghouse Electric Corporation, East Pittsburgh, Pa.): We are particularly interested in the predictions as to the future use of 400-cycle power by the Navy.

The minimum size and weight of the 600-kw 400-cycle 12,000 rpm, 4-pole cylindrical-rotor generator were achieved not only by an increase in speed and frequency, but by other means as well. Intensive design work resulted in the maximum reduction in the size of working parts, and the use of high-strength materials provided a reduction in the weight of such structural parts as the frame and bearing brackets. Furthermore, this design required the use of high-precision machining and manufacturing techniques. In this regard, we now consider that procurement and repair of such generators could be facilitated if less precise procedures were adopted. This would involve a slight increase in size and weight.

Construction of subsequent generators, using the *Timmerman* design as a prototype, demonstrates the effectiveness of the design as originally developed, although for these later generators it was necessary to make modifications where possible, to meet specification requirements which had not been imposed at the time the *Timmerman* design was prepared.

The development of the 12,000-rpm rotary amplifier exciter for use with our a-c generator on the *Timmerman* was noteworthy in demonstrating what can be accomplished in the design and manufacture of high-speed d-c machinery. The very high speed of the exciter resulted in critical commutation problems that required precise adjustment and mechanical balance. The delicate parts made necessary by its small size made it difficult to maintain the proper adjustments. Though the operation was successful with proper maintenance, later units used static exciters because they required less attention.

The authors' paper emphasizes that reduction of size and weight was the primary goal to be achieved by a change from 60 to 400 cycles. We believe that another important step toward the achievement of minimum size and weight for machines of any frequency could result from close scrutiny by the Navy of its specifications, so as to eliminate any requirements which are not actually needed for each application being considered. A typical example is the standard requirement for short-time overload capability, coupled with a special restriction as to maximum hot-spot temperature for all operating conditions, including overloads. Such a combination of requirements necessarily increases the size and weight of an electric machine, so it should be specified only when actually needed. Thus, a relaxation of any restriction which would result in oversize equipment will assist the Navy toward its goal of achieving minimum size and weight. We suggest that this phase of the problem be actively pursued.

P. L. Alger (General Electric Company, Schenectady, N. Y.): The 400-cycle power developments described have been most helpful in pushing along the progress of lower frequency equipment, as well. Designers of 60-cycle motors, challenged by these efforts and spurred on by the demands of industry, have made a great deal of progress since the *Timmerman* project began.

For example, the weight of the standard 40 C rise industrial motor, rated 30 hp, 1,200 rpm, of 1955 design, is 577 pounds instead of the 850 pounds given in Fig. 1. By going to 60 C rise, this weight comes down to 498 pounds. And, the comparable 60 C rise, 60-cycle 30-hp motors of 1,800 and 3,600 rpm weigh only 325 pounds and 290 pounds respectively. Furthermore, these industrial motors can all operate without injury at 115% of rated load continuously, and up to 150% for brief periods of time, whereas the higher temperature, 400-cycle motors have far less overload capacity.

In my opinion, it will be wise for naval vessels to use electric equipment that is only one or two steps removed from typical industrial designs, because of the great advantages in availability, reliability, versatility, and cost thereby secured. By so doing, existing production tools may be

used, giving immediate and high production capacity with low costs for naval as well as industrial motors. If such an extreme departure as 400 cycles is made, and if there is a sudden demand for large numbers of motors for Naval use, tremendous costs and delays will ensue.

Industry is steadily reducing the size and weight of electric motors, but so far only of machines of moderate temperature rise. This is because high-temperature motors have less starting, reversing, and overload ability, they are painful to touch, and offer little, if any, cost advantage. In industry also, direct-drive motors without gears are used wherever feasible. While the speeds of driven equipment have been gradually rising, only a very small proportion of industrial machines is adapted for speeds of over 3,600 rpm, and by far the greatest number are suited for speeds of only 1,800 and under. Therefore, if 400-cycle motors are used for power equipment, it will usually be necessary either to employ gears or use motors with more poles and consequently lower pf, higher weight, and higher cost. Hence, it has been found in industry, and the same is probably true on shipboard, that the most economical frequency is still 60 cycles for power installations larger than 1,000 kw or so.

As the authors indicate, the cables weigh one half as much as all the rest of the electric equipment on shipboard. Raising the frequency increases the voltage drop in the cables and makes more difficult the problems of motor-starting current and pf, voltage regulation, etc. Thus, nearly all the saving in cable weight, by going from 450 to 1,000 volts, and which would be available with 60 cycles, is lost at 400 cycles.

Nearly all the advantages obtainable with a 400-cycle power supply can be obtained with less cost and greater reliability by adopting a frequency of 420 or 180 cycles. If any departure from the usual 60-cycle frequency is made for general power applications on shipboard, the new frequency should certainly not be higher than 180. In my opinion, a 60-cycle supply is still the best, considering all factors.

William A. Hall (General Electric Company, Schenectady, N. Y.): This is quite an optimistic picture of the benefits derived from the use of a 400-cycle system on board a naval vessel. We agree with the prediction that, as far as certain specialized usages are concerned, there will be a considerable extension of the use of 400-cycle power on future vessels. The use of a 400-cycle supply for electronic equipment, aircraft servicing, and fluorescent-lighting installations does result in appreciable weight reductions. Perhaps in the future the use of the higher frequency can advantageously be extended to other more or less distinct areas such as, for example, the ship's ventilating system, where the number of different ratings of motors is small.

The difference in size and weight between the *Timmerman* components and those of a conventional destroyer is startling, and we are prone to overlook the real reasons for this great difference, and attribute it to the use of 400-cycles, as though this were a magical formula. Fundamentally, the 400-cycle system enables us to employ higher speeds. For a given hp, we need proportionately less torque at the higher speeds. In designing for less torque, we can reduce the physical size of all moving parts. Thus it can be seen that the frequency is instrumental in reducing size only in so far as it permits us to employ higher rotational speeds. Unless the driven auxiliary can effectively utilize the higher motor speeds, the weight reduction accomplished in the motor is at least partially offset by the necessity of adding reduction gears between the motor and the driven auxiliary. The major reason for the tremendous difference in size and weight of motors, for example, is that the *Timmerman* contract allowed the motor engineers complete freedom in design. Safety factors and design margins were held down to bare minimums. The resulting motors, while quite successful when judged as an experiment, are a far cry from being a suitable prototype design for a 400-cycle system. If the *Timmerman* motors could be redesigned to incorporate the long life and the reliability of the modern 60-cycle Navy motor there would be much less weight advantage in their favor.

However, 400-cycle motors for general application on board ship are a long way from being realities. They are, in fact, highly impractical except on an experimental ship such as the *Timmerman*. These motors differ too much from industrial 60-cycle motors to permit the building of both types with the same production facilities. In the event that even reasonable quantities of 400-cycle motors were required for a naval program, new facilities would have to be set up for their manufacture. In the event of an "all-out" defense effort all available motor manufacturing facilities would require drastic retooling and modification to enable them to produce 400-cycle motors, most existing test equipment would have to be scrapped, and new power supplies obtained.

There are other disadvantages: the lack of engineering talent possessed of 400-cycle experience, the dearth of machinists possessing the skills necessary for this very precise and accurate work, and the lack of naval repair facilities for repairing high-speed, accurately balanced motors. But these are all relatively minor factors when compared with the lack of facilities for large-scale production of the motors.

In the past, naval electrical machinery has been patterned quite largely after industry standards. As a result, in times of emergency, all of the industrial manufacturing capacity of the country has been available almost instantly for military production. This will not be the case in the future if the design of naval electric machinery is allowed to depart too radically from that employed in industry.

Lieut. C. E. Donaldson, III (U. S. Navy, Navy Yard, Brooklyn, N. Y.): My remarks deal primarily with the operation and maintenance aspects of the high-frequency power supply system of the Navy's experimental destroyer, *USS Timmerman*. Perhaps a word about the *Timmerman* design philosophy is in order. The *Timmerman* was not developed as a prototype vessel in any sense, but rather as a practical floating laboratory wherein the most futuristic and advanced designs in both marine steam propulsion equipment and auxiliary electric power systems could be installed in a standard destroyer hull and tested under seagoing conditions. It is significant that a great amount of purely laboratory-style instrumentation is built into this vessel. The original design requirements stipulated only that the resulting equipment meet the usual standards of quality and good workmanship, and that it further meet Navy high-impact shock standards. Indeed, the early implication was that if the designers produced equipment which functioned satisfactorily initially, this would constitute de facto evidence that the design had not been carried to the limits of weight and size reduction. The resultant engineering plant, developed by many companies, has spawned a great many practical equipments now in standard use in both the Navy and industry.

A few points should be specifically mentioned in amplification of the authors' paper. If the Navy eventually develops warships whose entire electric system operates at 400 cycles, the need for large-scale conversion of power from 60 to 400 cycles will arise, as, for example, during periods when such vessels might be idle alongside a pier at a shipyard under "cold-plant" conditions. The *Timmerman* shore power conversion equipment furnished approximately 418 cycles (2-pole 3,550-rpm 60-cycle induction motor direct-connected to a 14-pole 3,428.57-rpm 400-cycle generator). At this frequency, some motors were observed to draw as much as 110% to 115% of normal full-load current, due to the increased speed of rotation and the resultant increase in horsepower delivered to the connected load. Had the system been designed for 420 cycles this problem would not have arisen. Also, synchronous-synchronous conversion equipment might have been employed giving the added advantage of transferring the frequency stability of any large 60-cycle shore system to the 420-cycle system.

It is apparent that much room for improvement exists in the field of driven auxiliaries for high-speed motors. While hull ventilation fans and several liquid pumps were direct-connected to 8-, 6-, and 4-pole motors (6,000, 8,000, and 12,000 synchronous rpm), approximately 50% of the driven auxiliaries were driven at conventional speeds through gearmotors. Some development work was undertaken on 24,000 rpm rotary compressors for Freon-12 and water vapor, but no satisfactory design was produced. Recent developments in the aircraft industry might indicate methods of improvement for high-speed, compact refrigerant compressors, air compressors, and water pumps.

At the high-voltage, high-power levels involved, the 400-cycle switchgear operated satisfactorily except under fault or over-current conditions at low power factor. Quite often, arc interruption in the Navy *AQB* circuit breakers required as much as 0.50 to 0.75 second (200–300 cycles) and the action was not unlike that of d-c switchgear. It is postulated that the very short deionizing time within the arc chambers produces arc decay characteristics similar to those for d-c equipment.

After correction of some relatively minor difficulties, the *Timmerman* generating equipment was eminently quite satisfactory. The 24,000-rpm 750-kva synchronous turbogenerator must be seen to be appreci-

ated, since it represents just about the ultimate in size reduction.

Mechanically, after some corrective action, all the generators could be classified as very good. They are electrically sound, the static excitation systems function nicely and the magnetic-amplifier type of voltage regulators are outstanding, particularly in speed of response.

The biggest single problem in the high-speed motor program arose in the antifriction bearings. The combination of the high temperatures made possible by the use of class-*H* insulating materials and the high rotating speeds created a set of really severe bearing conditions. In some instances the speed-bore figure (speed in rpm times bore in millimeters) for certain bearings exceeded Anti-Friction Bearing Manufacturers Association recommended maximum figures, but this was just a manifestation of the original design philosophy previously mentioned. Careful choice of greases, with both organic and silicone bases, plus close attention to machining of motor components and assembly techniques resulted in marked improvement in performance.

The class-*H* insulation in the motors showed good moisture resistance and excellent heat resistance. Use of silicone-enamelled wire for rewinding effected a considerable saving in both slot space and production time over the original glass-insulated wire. Over 90% of all the motors were equipped with complete class-*H* insulation systems.

The high-frequency–high-voltage system offers interesting possibilities in the field of pf correction by the use of conventional askarel-filled paper capacitors. Several 0.5-microfarad units rated at 3,000 working volts direct current were successfully utilized directly on the 1,000-volt 400-cycle line giving a correction of 2.5-kvar per microfarad. Adequate ventilation is a necessity.

In conclusion, it may be mentioned that a 24,000-rpm 1,000-volt 600-kw capacitor-excited induction turbogenerator was developed for the *Timmerman* but was never installed on board.

J. F. Convery (Gibbs and Cox, Inc., New York, N. Y.): The authors' excellent paper is a concise summary of an important experiment. Ships are so expensive that experimental programs such as that conducted on the *Timmerman* are rarely undertaken, and usually only by a governmental agency. This may be unfortunate from the viewpoint of those who are impatient for progress at any cost, but is inevitable when the investment in any ship is considered. We in the industry are fortunate that the Navy had the foresight and enthusiasm necessary to push through the *Timmerman* project. Much of the work done there will eventually result in increased carrying capacity and earning ability of cargo and passenger vessels.

The paper states that of the total weight savings, 15% is attributable to modern design, 5% to the increase from 450 to 1,000 volts, and 10% to the increase from 60 to 400 cycles. A different way of considering the savings, based on studies made when the *Timmerman* was delivered, suggested that, considering the total weight saved as 100%:

35% was due to design advances which have since been utilized to some extent in shipboard equipment.
15% was due to design advances, which possibly were too advanced.
30% to 35% was due to the increase from from 60 to 400 cycles.
15% to 20% was due to the increase from 450 to 1,000 volts.

It is considered that one of the greatest benefits to the country and to industry of this program is implicit in the summary just given. Some participants were reluctant at first to devote engineers' time to developing equipment for the *Timmerman*. However, even casual inspection of these companies' recent designs shows evidence that their efforts were not wasted. Equipment now commercially available incorporates many features first used in the experimental destroyer. In fact, some of the companies, who needed the most persuasion at first, are now on their own initiative proposing design programs of several types.

J. M. Apple and E. W. Lusby: The authors wish to thank the discussers for their interesting comments and constructive information. As can be seen, there is no unanimity on the use of 400-cycle power. However, the authors would like to point out that if the use of 400-cycle power does increase in the future, it will probably be because of the necessity for meeting the needs of the power users, rather than for any needs of the power generation and distribution systems. The following are answers to the specific questions raised in the discussions.

Mr. Shuster requested information pertaining to the use of direct-driven auxiliaries, tests on the induction generator, and pf correction. On the *USS Timmerman* approximately 55% of the total motor-driven auxiliary hp rating, other than fans, are direct-driven. The induction generator, built in conjunction with the *USS Timmerman*, was tested but voltage regulation problems were not satisfactorily solved. The 24,000-rpm direct turbine-driven generator performed exceptionally well, but considerable difficulty was encountered with the capacitor-reactor voltage regulator-exciter system. Capacitors were provided at the motor controller in some instances for pf correction. Owing to numerous capacitor failures this correction was eventually removed and oil-filled capacitors were provided at central points in the distribution system for this purpose.

Mr. Gatewood's questions are answered in the order presented.

1. Navy experience has indicated that selenium rectifiers are just as practical with a 400-cycle supply as with a 60-cycle supply. Most of the difficulty with selenium rectifiers is attributable to misapplication, particularly insufficient derating of the rectifiers for the temperature involved in the application.
2. Across-the-line starting is practical on 400-cycle systems and was exclusively used on *USS Timmerman*.
3. Type *SGA* Navy cable is rated for 1,000 volts.
4. Cables were derated for 400-cycle application, as mentioned in the paper.
5. The authors' intent in the paper was to stress the importance of alignment, fits, and balancing. Methods for accomplishing this on electric motors should be reasonably simple and capable of being accomplished on shipboard.

Mr. Mayer's inquiry as to what extent synchronous speeds of 6,000 to 8,000 rpm influenced the characteristics of either the driven auxiliaries or associated systems is quite adequately answered by the ventilation system provided on the *USS Timmerman*. Running at a synchronous speed of 8,000 rpm, it was found possible to make the fans and vent ducts of the complete ventilation system considerably smaller.

Sealed Ignitron Principles Extended to Larger Tubes

J. L. BOYER
MEMBER AIEE

A. P. COLAIACO
ASSOCIATE MEMBER AIEE

THE universal acceptance of the sealed ignitron for industrial power rectifiers up to 1,000 kw at 600 volts has raised the question as to the possibility of larger sealed ignitrons which would permit higher ratings for rectifier units. After considerable development, the same principles used for sealed ignitrons up to 400 amperes have been extended to larger sealed tubes with ratings up to 1,000 amperes at 250 volts d-c.

In the past, the higher power ratings could be obtained by using many small tubes in parallel, but this was not always economical because of the complications involving the power and excitation circuits. Now that large, sealed tubes are available, this condition is completely changed and good designs of sealed ignitron rectifier units can be obtained with ratings up to 2,000 kw for six tubes at 600 volts. Even larger installations can be obtained by operating these large, sealed ignitron rectifiers in parallel.

The developments of two new large sealed ignitrons were based on the design and manufacturing principles which have been established on thousands of sealed tubes over a period of 17 years. These principles provide rather tight limitations on what must be done so as to obtain reliable performance and long tube life. Many variations of these principles have been studied and some offer short-range economics but, in the final analysis, it has been found that the highest quality sealed ignitrons must be based on these principles.

Principles of Design of Sealed Ignitrons

The development of larger sealed ignitrons require that certain principles, which have been proved by many years of manufacture on smaller tubes, be applied to the electrical and mechanical designs. The electrical designs are based on principles which are common to all types of mercury-arc rectifiers, and these have been completely covered in previous papers.[1,2] The mechanical designs are based on principles which are special for sealed ignitrons, where the major concern is the problem of keeping a low gas pressure inside the tube for a long period of time under all conditions of operation.

The following principles have been found to be of greatest importance in the manufacture of sealed ignitrons:

1. All insulating bushings and seals must be vacuum-tight and have sufficient mechanical strength under all degassing and operating temperature, and must not be subject to deterioration with time.

2. All exposed materials inside the tube must not react with ionized mercury vapor.

3. Stainless steel or other corrosion-resisting materials must be used for the water system and main cylinder.

4. All joints must be welded with materials which have high strength at the tube degassing temperature.

5. Anodes and grids must be designed so that absorbed gases in the graphite will have a minimum distance to diffuse out during degassing.

6. Graphite electrodes must be given a first degassing for a reasonable period of time at a temperature far above any they may experience in normal operation.

7. The walls and major surface of the tube must be degassed at the highest practical temperature so that its gas-absorbing capacity during operation is at maximum.

8. Graphite electrodes must be given a final degassing inside the tube at about the same temperature experienced in normal operation. This degassing must be done at a time when the tube wall also is being degassed.

9. Over-all tightness of the completed tube must be such that no leak can be detected with the present commercial mass spectrometer leak detectors.

10. Materials and processes must be such that over an extremely long operating time, the formation of mercury sludge is low. The design also must be such that sludge has the minimum effect on operation.

The principles 1 through 5 mean that the sealed ignitron must be designed and made from materials which will ensure a vacuum-tight tube for all service conditions. This requires the use of materials which can be operated or subjected to very high temperatures without mechanical failure or chemical change. In practice, it is found that sealed ignitrons can be designed and manufactured so that these principles are completely satisfied and, as a result, practically no tubes are ever lost because of vacuum leaks.

The principles 6 through 8 mean that the sealed ignitron must be put through manufacturing steps which thoroughly degas all parts, so that a minimum of gas is released during normal operation and the tube wall and action of the arc will absorb any gas released because of overload or fault conditions. The thoroughness with which the principles are applied determines the life of the majority of sealed ignitrons. This is because the most important factor which determines tube life is the amount of gas released during abnormal operation and the effectiveness with which it is absorbed in normal operation.[3]

The principles 9 and 10 are over-all measures of how effective the materials, design, and degassing are in making a vacuum-tight tube which will not be contaminated by impurities in any form. In determining ignitron life, these last two principles are measured in terms of what effect they have on tube performance in arc-back rate and misfire rate.

Basic Construction of Large, Sealed Ignitrons

The design principles of the new 667-ampere and the 1,000-ampere sealed tubes are based on the experience obtained with several sizes of smaller sealed tubes of which thousands are in operation. These tubes have given many years of satisfactory service, and provide an excellent basis for evaluating the performance to be expected from the new and larger sealed tubes.

Figs. 1 and 2 show the close similarity between the new 667-ampere and the next smaller ignitron, the long-established 400-ampere tube, some of which have been in continuous operation for over 8 years. The tube walls and water jacket are made of the same columbium-stabilized, stainless steel to resist corrosion. Both the anode and cathode headers can be removed for reconditioning the tube by grinding off the arc welds. All insulating seals are of the glass-Kovar type, which have proved to be both rugged and vacuum-tight. The internal graphite electrodes are predegassed in a vacuum furnace at 1,800 degrees centigrade (C) before being assembled into the tube.

Paper 56-173, recommended by the AIEE Industrial Power Rectifiers Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Winter General Meeting, New York, N. Y., January 30–February 3, 1956. Manuscript submitted October 24, 1955; made available for printing December 1, 1955.

J. L. BOYER and A. P. COLAIACO are with the Westinghouse Electric Corporation, East Pittsburgh, Pa.

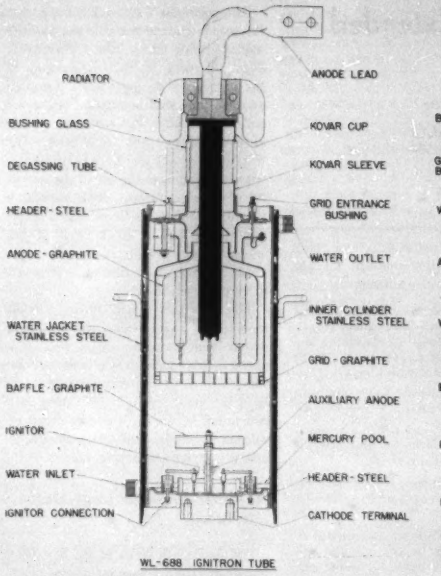


Fig. 1. Cross section of 400-ampere ignitron tube

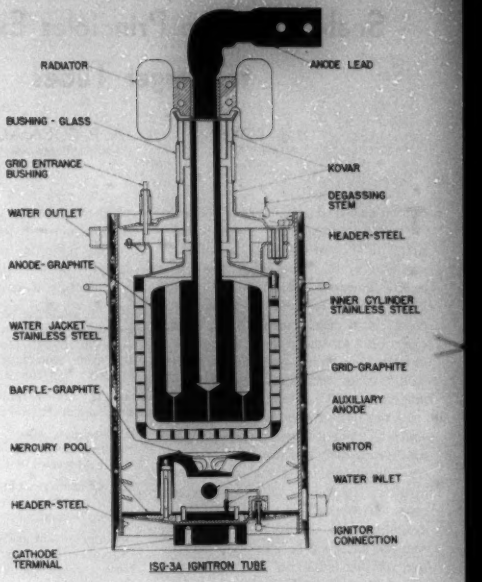


Fig. 2. Cross section of 667-ampere ignitron tube

After assembly and before the mercury is introduced, the tube is degassed for many hours in a furnace, with the main body of the tube being heated to 900 C. It is important to note that the anode temperature is approximately 900 C during this degassing period. In fact the tube degassing temperature is determined by the normal maximum operating temperature of the anode. It is this manufacturing operation which provides the final degassing of the anode without putting the gas into the tube walls and exposing the operating mercury vapor arc to the gas released by the anode.

After degassing, the tube is checked for vacuum tightness with a helium leak detector of the mass-spectrometer type. The sensitivity of this leak detector is such that it will detect a leak which would raise the pressure in the tube less than 10 microns in 10 years. Thus, leakage of external air into the tube will have a negligible effect upon its operation and does not determine tube life. The final seal is made by pinching off the degassing stem with a resistance welder. Before shipment the tubes are given a high-potential test to determine whether the tubes have been damaged in handling.

The 1,000-ampere sealed tube is constructed similar to the 400-ampere and 667-ampere tubes except for some minor modifications made necessary because of its larger size. The tube wall and cathode are made of one piece, drawn-tank construction with the cathode as well as the side wall being water-cooled. The anode header is removable for reconditioning the tube. Because of the large, heavy anode used in this tube, the main glass-Kovar serves as a vacuum seal and does not support any weight. A porcelain bushing entirely surrounds the glass-Kovar seal and carries the main weight of the anode. Thus, the main bushing assembly has the vacuum tightness of glass-Kovar but with the added strength of porcelain.

Ratings of Sealed Ignitrons

Table I shows the ratings of a series of sealed ignitron tubes for industrial service including the two new tubes. Fig. 3 shows all these tubes. It is apparent from the table that the maximum kilowatts available in 6-tube and 12-tube assemblies have been more than doubled thus affording a wider choice in the selection of sealed ignitron rectifiers.

The average current rating of a sealed or pumped ignitron is approximately proportional to the internal diameter of the tube. Since the volume of the ignitron is about proportional to the square of the diameter, the volume of the tube per rated ampere goes up as the square of the diameter. Fig. 4 shows the relationship between the average current rating and the net tube volume. For the larger current ratings, it can be seen that the volume per ampere is much greater for the larger tubes. As a result, pressure rises in the larger sealed ignitrons are not as great when there is a fault or overload of the same relative magnitude. This characteristic tends to give longer useful life to the larger sealed ignitrons.

Table I. Sealed Tube Current Ratings

Tube Size	Maximum Amperes: Average Current	2 Hours	1 Minute	Surge Current (0.015 Second)
		230 Volts D-C		
C	100	150	200	6,000
D	200	300	400	12,000
E	400	600	890	25,000
G	667	1,000	1,333	40,000
K	1,000	1,500	2,000	60,000
		600 Volts D-C		
C	75	112	150	4,500
D	150	225	300	9,000
E	300	450	600	19,000
G	415	622	830	25,000
K	610	915	1,220	45,000

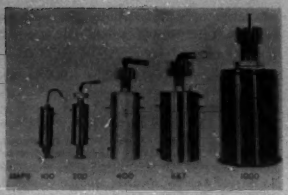


Fig. 3. Series of sealed ignitron tubes

Sealed Tube Life Expectancy

In the practical operation of sealed ignitrons, many conditions influence tube life, and to some extent they can be controlled. Tube life is dependent primarily on the following factors:

1. The quality of the tube design and manufacture.
2. The actual loading on the tube compared to the tube rating.
3. The frequency and magnitude of overloads and system faults.
4. The performance of the protective equipment.
5. The care taken in maintaining the cooling system.

For a given tube design, factors 2 through 5 may vary considerably, making it difficult to predict tube life and compare tube life between different installations. An indication of life expectancy may be obtained from Fig. 5. These curves show the survival rate for three different sizes of tubes. The curve for the 667-ampere tubes extends only to 12 months, as this is the operational experience to date on a reasonable number of tubes.

A survey was made of the 30 original 400-ampere tubes installed in 1947. This survey showed that after 8 years of operation, 18 of the original tubes, or 60 per cent, were still in operation. Since these first 30 tubes were built, numerous improvements have been made in the design and manufacturing procedures of the tubes so that an even higher survival rate is obtained on the more recent tubes. There have been numerous cases where sealed tubes have been in continuous operation for 15 years.

Fig. 5 also indicates that larger tubes have a higher survival percentage. These data confirm the fact that longer life is obtained on larger tubes since the volume goes up with the square of the diameter, whereas the tube rating only goes up linearly with the diameter.

Sealed Ignitrons Versus Pumped Ignitrons

The development of large, sealed ignitrons has made it possible to design larger sealed rectifiers which can be applied to the same ratings as pumped ignitrons. As a result, it is necessary to decide whether a sealed or pumped rectifier should be used for a particular application. The decision must usually be based on economic factors but, at times, there are other conditions which govern the choice.

The selection of a sealed ignitron rectifier versus a pumped ignitron rectifier for most applications can be determined by weighing the cost of repair, or replacement, of the sealed ignitron against the cost of maintenance of the vacuum system of the pumped rectifier. The average annual cost of sealed tube replacement is extremely low, since sealed ignitrons in most applications give very good life. However, a recent AIEE survey indicates that, for pumped ignitrons, the number of maintenance hours attributable to the vacuum system is only 12 per cent of the total equipment maintenance.[4] Thus, the economic factors are not large, and the application of sealed or pumped ignitrons may be determined by whether maintenance personnel will be available at the installation or not. If the rectifier unit is to be located where there will be no maintenance personnel, the rectifier should use sealed ignitrons.

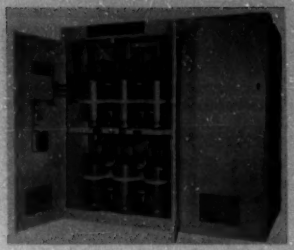


Fig. 6. Cubicle of six 667-ampere ignitron tubes

If maintenance personnel will be available for other purposes, the rectifier should use pumped ignitrons.

The end result of these considerations is usually that sealed ignitron rectifiers are applied in the smaller installations where only several rectifiers are required, and that pumped ignitron rectifiers are applied in the larger installations where many rectifiers are required. Some exceptions to this rule can be justified where special conditions make it desirable to use one type of ignitron over the other. However, unless there are good reasons, some care should be taken so as not to pay both the cost of replacing or repairing sealed ignitrons, and the cost of maintenance personnel.

Fig. 6 shows six 667-ampere tubes in a metal-enclosed cubicle with an industrial rating of 1,000 kw at 250 volts. The excitation circuits are located directly in back of the tubes. The right-hand compartment encloses a water-to-water heat exchanger.

Conclusions

Sealed ignitrons designed and manufactured by the principles indicated have been found to have an average life greater than 8 years. The loss of sealed ignitrons caused by vacuum leaks is a very minor problem and for all practical pur-

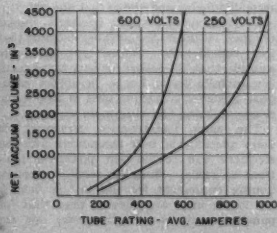
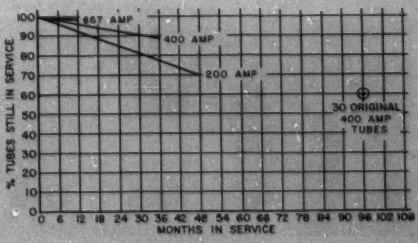


Fig. 4 (left). Tube ratings versus net vacuum volume

Fig. 5 (right). Life expectancy of sealed ignitrons



poses almost no tubes fail for this reason. The major cause of limited tube life is the rise in pressure caused by foreign gases given off from internal parts being heated above the temperatures at which they were degassed. Therefore, the highest quality sealed ignitrons are obtained when the internal parts are very thoroughly degassed at the highest possible temperature and the tube walls are able to absorb the maximum of gas.

In practical operation of sealed ignitrons, many factors determine tube life, and it is only possible to define life in terms of typical averages for an application. For one tube design, the maintenance of the water-cooling system, the number of overloads, the minimum time between overloads, and the performance of the protective equipment greatly influence the average life in a given application.

The two large, sealed ignitrons, which have been designed and manufactured according to the principles used on smaller sealed ignitrons, are being applied to extend the ratings of sealed ignitron rectifiers. Experience indicates that, because of the greater volume per rated ampere, the larger sealed ignitrons have a longer average life than smaller tubes.

Sealed ignitron rectifiers of the larger ratings should be used in applications where maintenance personnel are not available. This is usually the case in small installations or where only a small number of rectifiers are to be installed.

References

1. Mercury Arc Rectifiers and Ignitrons, J. H. Cox, D. E. Marshall. *Transactions*, Electrochemical Society, New York, N. Y., vol. 72, p. 183.

2. A 400-Ampere Sealed Ignitron, H. C. Steiner, H. N. Price. *AIEE Transactions (Electrical Engineering)*, vol. 65, Oct. 1946, pp. 680–85.

3. The Absorption of Gases by the Steel Walls of Vacuum Discharge Apparatus and Its Bearing on the Life of Pumpless Mutators, W. Deck. *Brown Boveri Review*, Baden, Switzerland, vol. 29, no. 8, 1942, pp. 202–04.

4. Survey of Operations of Mercury-Arc Rectifiers, AIEE Committee Report. *AIEE Special Publication S-47*, Apr. 1952.

Discussion

L. W. Morton (General Electric Company, Schenectady, N. Y.): This paper discloses the good news that the authors' well-established series of sizes of smaller sealed ignitron tubes has been extended by the addition of two new larger tubes. The ratings of these new tubes match the present ratings of large pumped ignitron tubes. Extension of sealed ignitron principles to these larger ratings by a second group of engineers is gratifying, for it tends to confirm the soundness of a trend of several years toward larger sealed ignitrons.

The group of engineers with whom I have been associated completed the development of a sealed tube having the same rating as the authors' size *G* tube in 1948, and completed development of one having the same rating as their size *K* tube in 1951. E. J. Remscheid's conference paper entitled, "Service Life of Pumpless Ignitrons," presented this year, summarizes the extensive field experience acquired with these large sealed tubes since their introduction. On the basis of his report, the future looks bright for large sealed ignitron tubes.

The developmental approach of the group of engineers with whom I am associated was different from that of the authors' group. The latter extended the design of their smaller sealed ignitrons to the larger sizes. We simply converted our large pumped ignitron tubes into large sealed ignitron tubes. A glance at Fig. 3 shows the striking family resemblance of all of the authors' tubes, both large and small. A similar comparison of our series shows a dissimilarity between the two largest and the smaller sizes; see Fig. 7. The ratings of the authors' series of sealed tubes correspond to those of ours.

In 1948, we chose to convert our pumped tubes to the sealed type for practical reasons. Our pumped tubes were well developed and time-proved in service. The developmental problem was limited to finding and applying techniques for making these large tubes as vacuum-tight as necessary, and to degassing them thoroughly. Not all of the techniques used with the smaller sealed tubes, developed several years earlier, were directly applicable but they helped. We soon learned to ensure vacuum tightness with the mass-spectrometer leak detector and to obtain thorough degassing by somewhat different methods from those used on the smaller sealed tubes.

We were spared the need to retrace all our factory and field rating tests because the new large sealed tubes were still the same in design as our proved pumped types. Enough check points were tested to establish that the tubes performed as well sealed as pumped, after which they were ready for use.

A few improvements have been made during the 7 years since we put our first large sealed rectifier in service. Several years ago, the vacuum chambers and cooling jackets were changed from ordinary steel to stainless steel to eliminate the possibility of corrosion or hydrogen ion penetration. A spare ignitor was added at the same time to avoid the need of opening the tube in case of failure of one of them. Since much of the favorable field experience reported by Mr. Remscheid involved the original pumpless tubes built of ordinary steel and having only one ignitor, we may look forward to even better field experience from our pumpless tubes of improved design. The improvements just described did not affect the ratings or performance. Many of the parts of our large sealed ignitrons are still interchangeable with those of the pumped type, and the ratings are still identical. The performance of all of our large sealed tubes has been, if anything, superior to that of the corresponding pumped type.

Perhaps a few remarks about the significance of the terms "sealed tube" and "pumpless" rectifier are in order. A pumpless rectifier, is, of course, a sealed tube rectifier. Now, why have the engineers with whom I am associated used the name "pumpless?" L. A. Umansky, this year's recipient of the Edison Medal, first coined the term "pumpless-pumpable" to describe the kind of sealed tube which we were developing. The term was appropriate because our objective was to develop a reconditionable sealed tube. The tubes of smaller size already in use at that time were expendable; when they failed, it was less expensive to replace them than repair them. For the larger sized tubes, it was obvious that it should be less expensive to recondition them than to replace them. The longer term "pumpless-pumpable" was quickly shortened to "pumpless," and we have continued to use the term, chiefly to distinguish between the sealed tubes that are reconditionable and those that are not.

In comparing sealed ignitrons versus pumped rectifiers, the authors indicate that they consider the pumped type still more suitable than the sealed for very large rectifier installations. Large pumpless rectifiers, however, are being used in very large electrochemical and aluminum plants.

The same maintenance personnel, in users' plants, who now recondition pumped rectifiers can recondition pumpless ignitrons as easily. Pumpless rectifier reconditioning equipment for use in the field is not much more complicated than that for pumped rectifiers. Furthermore, our ex-

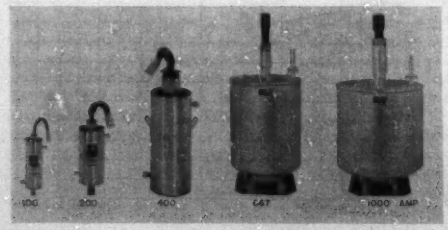


Fig. 7. General Electric Company sealed ignitron tubes, showing different sizes

perience indicates that the pumpless type will require less frequent reconditioning than the pumped.

We believe that in time the new large sealed tubes which have been so well described by the authors will prove more suitable even for the largest rectifier installations than the pumped type.

C. C. Herskind (General Electric Company, Schenectady, N. Y.): This interesting paper confirms the strong trend toward the elimination of pumps on large mercury-arc rectifiers. The sealed ignitrons described in the paper cover the same field as the pumpless ignitrons introduced 6 years ago. On the question of nomenclature, the terms "sealed" and "pumpless," seem to be synonymous.

The development of sealed or pumpless tubes in the larger sizes has been slow because of the lack of outstanding economic advantages. However, the acceptance of the pumpless rectifier and the development of these large-sized sealed ignitrons indicates that users find other important advantages. One of these is the reduction in specialized personnel through the elimination of pumping equipment and degassing operations in the field. Pumpless rectifiers eliminate the occasional need for user personnel having the special skill required for vacuum work. Another advantage is the safety to personnel resulting from the metal-enclosed construction. Perhaps users can point out others.

The paper shows the relation between tube ratings and net vacuum volume in Fig. 4 but the authors do not supply data indicating that this is a fundamental factor. Tests on pumpless rectifiers have not indicated that volume is important.

To obtain more information about the clean-up process, we have been making tests on a pumpless rectifier, operating a rectifier tank under load at normal current and voltage, admitting a known quantity of gas, and observing the clean-up action and arc-back performance. Similar tests under load have also been made admitting air continuously at a high rate. These tests have demonstrated a powerful clean-up action. They also show that clean-up is selective, with both the rate of clean-up and the residual pressure depending upon the kind of gas. Different gases affect operation differently, some apparently having little influence on arc back, while others cause arc back when present even in small quantities. We believe that a working knowledge of these facts is essential to the further improvement of high-power rectifiers.

J. L. Boyer and A. P. Colaiaco: The main purpose of this paper was to define and emphasize the principles which must be followed in the design and manufacture of sealed ignitrons. Mr. Morton's interesting discussion was greatly appreciated, but to some extent it indicated that the original paper was not definite enough to accomplish one of its objectives, which was to describe exactly the sealed ignitron so that the differences between sealed ignitrons and pumpless ignitrons could be distinguished.

Mr. Morton's position on pumpless ignitrons is rather surprising since the term "pumpless rectifier" has been used for many years with a definite meaning which is not the same as that of a "sealed rectifier."[1,2] This difference was recognized in the first paper presented by Mr. Morton's associates, "Development of a Pumpless Ignitron."[3]

Over a period of many years the sealed ignitron has been known as a tube which is completely hermetically sealed with no gaskets or porous materials used in the envelope and which is degassed at a very high temperature. A pumpless ignitron has been known as a tube which does not require vacuum pumps, which can be constructed with gaskets or seals with low leakage rates, and which is degassed at relatively low temperatures. Many designs of both these types of ignitrons have been made so that they can be reconditioned, and it is only correct to say that both sealed and pumpless ignitrons can be opened and reconditioned. An example of a sealed ignitron which can be reconditioned is the 400-ampere ignitron in Fig. 7, just as the 400-, 667-, and 1,000-ampere sealed ignitrons in Fig. 3 can be reconditioned.

It appears that Mr. Morton and his associates originally developed a true pumpless ignitron for the larger sizes of tubes and called it a pumpless tube. However, in recent years they have gradually changed the design and materials until the designs are now approaching true sealed tubes, and have decided to call them sealed ignitrons. This, of course, is a reasonable step and it emphasizes the contention of the paper that the sealed ignitron is the correct type for long tube life.

Mr. Morton indicated that it was very desirable to take internal designs proved on pumped ignitrons and use them on pumpless ignitrons. This is exactly what was done in the development of the larger sealed ignitrons described in this paper. The fact that the external appearance of the new sealed tubes is different from pumped tubes has nothing to do with the internal design. The sealed ignitrons have the same external appearance, since all of the tubes are based on the principles of construction proved best for sealed ignitrons.

Mr. Morton stated that sealed ignitrons can be reconditioned just as easily as pumped tubes. This is true if all that is done to the tube is to re-evacuate it and seal it off again. Even so, a vacuum system will be needed. It is apparent that just re-evacuating a pumpless or sealed ignitron with only a current degas is of very little value, since the tube walls are not degassed and there are still many solids in the mercury pool. It is usually desirable to recondition more completely the tube, and to open it and repair or replace internal components, especially the mercury. To do this requires considerably more equipment, such as degassing furnaces, welders, and even a mass-spectrometer leak detector is desirable. This represents a much larger investment than the few simple tools needed to recondition a pumped tube and requires more experienced personnel to do the work.

The discussion by Mr. Herskind raises some of the same questions on nomenclature, but it indicates a different point of view. He indicates that sealed and pumpless ignitrons are not exactly the same, since they are processed somewhat differently and the pumpless tubes required extensive tests to show that clean-up action is selective with both the rate of clean-up and the residual pressure depending upon the kind of gas. With sealed ignitrons this clean-up action has proved very satisfactory on a large number of tubes over a period of 17 years.

As to what measures are most effective in degassing, it can be stated that heat, vacuum, and time are the most effective. By raising the degassing temperature and degassing all parts in one process it is possible to obtain a much better degassed tube in the same length of time. The criterion of adequacy of degassing is the amount of pressure rise during normal and abnormal operation and the life of the ignitrons in actual service. The gas is removed by the clean-up action by driving the gas molecules into the tube walls and forming solid compounds.

Mr. Herskind states that tests on pumpless rectifiers have not indicated any relation between tube rating and net vacuum volume. There is probably no fundamental relationship but there is a resultant relation because of other characteristics. If the actual net volumes of any ignitrons are plotted against tube rating it will be found that the volume increases at a much faster rate than the rating. This principle has been confirmed by the laboratory and field experience of a large number of tube designs from many manufacturers. The higher ratio of net vacuum volume to tube rating for the larger sealed tubes explains at least in part why the larger tubes require less frequent reconditioning.

The question of whether sealed ignitrons or pumped ignitrons should be used in a given application is difficult to answer with a simple rule, since there are many conditions which must be considered. The paper presented the basic engineering facts to be considered and every effort was made to give a fair comparison even though the paper was describing sealed ignitrons. We believe that the future of sealed ignitrons can best be served by making applications only where the over-all characteristics and economics of the sealed tube are best for a particular application.

References

1. Sealed Metal Ignitrons. *Electric Journal*, E. Pittsburgh, Pa., vol. 34, Jan. 1937, p. 28.

2. Pumpless Steel-Bulb Mercury-Arc Rectifier, J. E. Boul. *English Electric Journal*, London, England, vol. 10, no. 6, Dec. 1943, pp. 144–47.

3. Development of a Pumpless Ignitron, C. C. Herskind, E. J. Remscheid. *AIEE Transactions*, vol. 70, pt. I, 1951, pp. 1066–69.

Block Diagram Transformations for Systems with One Nonlinear Element

T. M. STOUT
ASSOCIATE MEMBER AIEE

Synopsis: Rules for the transformation of block diagrams representing linear systems are examined to determine their applicability for systems containing nonlinear elements. General proofs and examples are given which show that the original block diagram for any system with one input, one output, and one nonlinear element characterized by $y=\varphi(x)$, where x and y are its instantaneous input and output, can always be reduced to four topologically identical and mathematically equivalent block diagrams. In these diagrams, the nonlinear element may appear as either a forward or feedback block with its input and output in either a normal or reversed cause-effect relationship.

BLOCK diagrams are widely used in the study of complex dynamic systems. In constructing these diagrams, mathematical relations between the system variables are shown by summing points which indicate additions and subtractions of the variables, and by blocks which express operations on the variables. For completely linear systems, the blocks indicate only differentiation or integration of the input signal with respect to time, multiplication of the input signal by a constant, or combinations of these operations; transformation theorems can be used to reduce the original diagram to an equivalent form.[1-6]

In systems which are not linear, one or more of the blocks will represent a nonlinear relationship between two or more variables, such as the flux-current relation of an iron-core inductor or the torque-flux-current relation of a d-c motor. Although the original diagram for a nonlinear system can always be drawn, there is no reason to expect that extensive reduction of the diagram will be possible. In this paper the applicability of the various transformation theorems to nonlinear systems is examined and the process of block diagram reduction for a restricted class of nonlinear systems is considered.

The nonlinear systems to be considered here are those systems containing one nonlinear element which can be characterized by a relation between the instantaneous values of its input and output, say $y=\varphi(x)$. Relations such as $z=\varphi(x, y)$, including $z=xy$ as a special case, are specifically excluded. Consideration of nonlinearities with a single input variable is less restrictive than it appears, although still a limitation. For example, a nonlinear inductor can be described by

$$e=N\frac{d}{dt}[\phi(i)]$$

instead of

$$e=[L(i)]\left[\frac{di}{dt}\right]$$

as is sometimes done. It would likewise be desirable to deal with systems having more than one nonlinear element, but it is doubtful if this could be done with any degree of generality; it can only be said that a significant number of the problems considered in practice involve just one nonlinear element.

It will be shown that the original block diagram for any nonlinear system of the assumed type can always be reduced to the standard forms shown in Fig. 1. In these diagrams, linear dynamic relations are indicated by the transfer functions G_1, G_2, G_3, and H, while the nonlinear relations are expressed by $y=\varphi(x)$ and $x=\psi(y)$. As will be pointed out, these diagrams are interchangeable descriptions of the same system if $G=1/H$ and ψ is the function which is the inverse of φ. These diagrams are generalizations of somewhat similar diagrams encountered in the work of Kochenburger[4] and Grief.[7] Although the paper deals directly only with electric circuits and feedback control systems, the results are clearly applicable to a variety of analogous systems.

Signal flow diagrams,[8,9] employing a slightly different notation and terminology, could be used in a similar way to obtain the same results. The standard block diagrams given here are, in fact, closely related to a diagram in the book by Truxal[9] (Fig. 2.29, p. 114); his discussion is, however, concerned only with completely linear systems. Block diagrams are used herein because it is assumed that engineers are more generally familiar with them.

Rules for Block Diagram Transformation

Graybeal listed[3] 21 rules for the transformation of block diagrams, 8 of which suffice for linear network analysis.[6] These eight rules, plus another rule not included in Graybeal's list but important in the present work, will be stated and examined for their validity for nonlinear systems.

The eight rules are shown pictorially in Fig. 2. The symbol A, with occasional subscripts, denotes a general operator

$$y=Ax$$

indicating that A, whether linear or nonlinear, operates on x to produce y. In writing equations

$$y=A\,.\,x \tag{1}$$

will denote ordinary multiplication by a linear operator, while

$$y=A(x) \tag{2}$$

will indicate that y is a nonlinear function of x. The symbol A^{-1} denotes an inverse operator such that $A^{-1}A\ x=x$. For linear operations, A^{-1} is simply the reciprocal of A; the inverse of a nonlinear operator is obtained by viewing x as a function of y instead of y as a function of x. The nature of A^{-1} is of no concern for system representation by means of block diagrams but may become important for system analysis. It is assumed that nonlinear relations concern only instantaneous magnitudes; any dynamic effects (integration, differentiation, etc.) are shown separately.

Rule 1 is valid only for linear systems where it represents

$$\begin{aligned} y&=A\,.\,x=A\,.\,(x_1\pm x_2)\\ &=A\,.\,x_1\pm A\,.\,x_2 \end{aligned} \tag{3}$$

since

$$A(x_1\pm x_2)\pm A(x_1)\pm A(x_2) \tag{4}$$

Rule 1 is, in fact, a definition of a linear

Paper 56-196, recommended by the AIEE Feedback Control Systems Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Winter General Meeting, New York, N. Y., January 30-February 3, 1956. Manuscript submitted October 13, 1955; made available for printing December 14, 1955.

T. M. STOUT is with the Ramo-Wooldridge Corporation, Los Angeles, Calif.

This paper is based in part on work done at the University of Washington, Seattle, Wash., under contract with the Office of Naval Research, and is included in a thesis submitted to the University of Michigan, Ann Arbor, Mich., in partial fulfillment of requirements for the degree for Doctor of Philosophy.

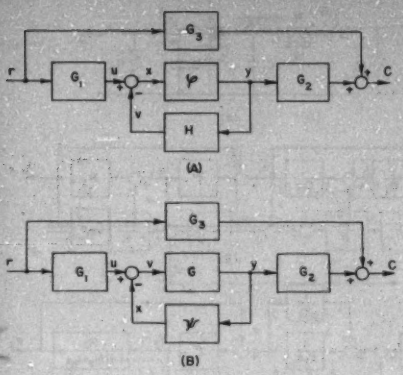
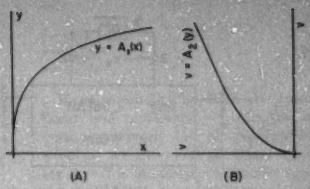


Fig. 1 (left). Standard block diagrams for systems with one nonlinear element



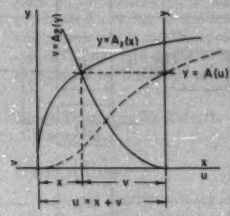


Fig. 3 (right). Graphical construction for nonlinear feedback system

operator and a statement of the superposition principle which applies only to linear systems. In similar fashion, rule 2 is valid only for linear systems; in combination, the first two rules state that linear blocks can be moved freely past summing points in either direction, but nonlinear blocks cannot.

Rules 3 and 4 state that blocks can be moved past take-off points in either direction and are applicable to both linear and nonlinear systems if proper consideration is given to the meaning of A^{-1}.

The equivalent operator required in rule 5 is defined by

$$z = A \,.\, x = [A_1 \,.\, A_2] \,.\, x \tag{5}$$

if A_1 and A_2 are linear, and

$$z = A(x) = A_2[A_1(x)] \tag{6}$$

if A_1 and A_2 are both nonlinear. A mixture of linear and nonlinear blocks cannot be combined unless the linear block represents multiplication by a constant; constants may be introduced into, or removed from, the description of nonlinearities to facilitate computations. It should be noted that, although the order of linear operations is immaterial, the order of a mixture of linear and nonlinear blocks is of vital importance and cannot in general be altered. For example, a differentiator followed by a limiter is not the same as a limiter followed by a differentiator.

Rule 6 is applicable to cases in which A_1 and A_2 are both linear, where

$$y = A_1 \,.\, x \pm A_2 \,.\, x = (A_1 \pm A_2) \,.\, x \tag{7}$$

or both nonlinear, where

$$y = A_1(x) + A_2(x) = A(x) \tag{8}$$

For a mixture of linear and nonlinear operators, an equivalent operator cannot be defined (unless the linear operator is a constant), but

$$y = (A_1 \pm A_2)x \tag{9}$$

can be used to signify that

$$y = A_1 \,.\, x + A_2(x) \tag{10}$$

if A_1 is linear and A_2 is nonlinear; in this case, the operators cannot be combined and the block obeys the rules for nonlinear blocks.

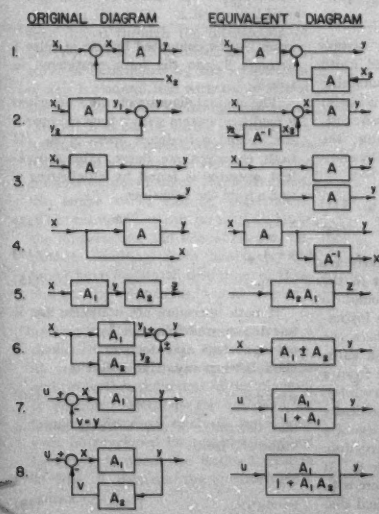
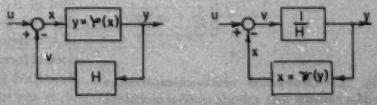


Fig. 2 (left). Rules for block diagram transformation

Fig. 4(A) (right). Block diagram transformation, rule 9. (B). An application



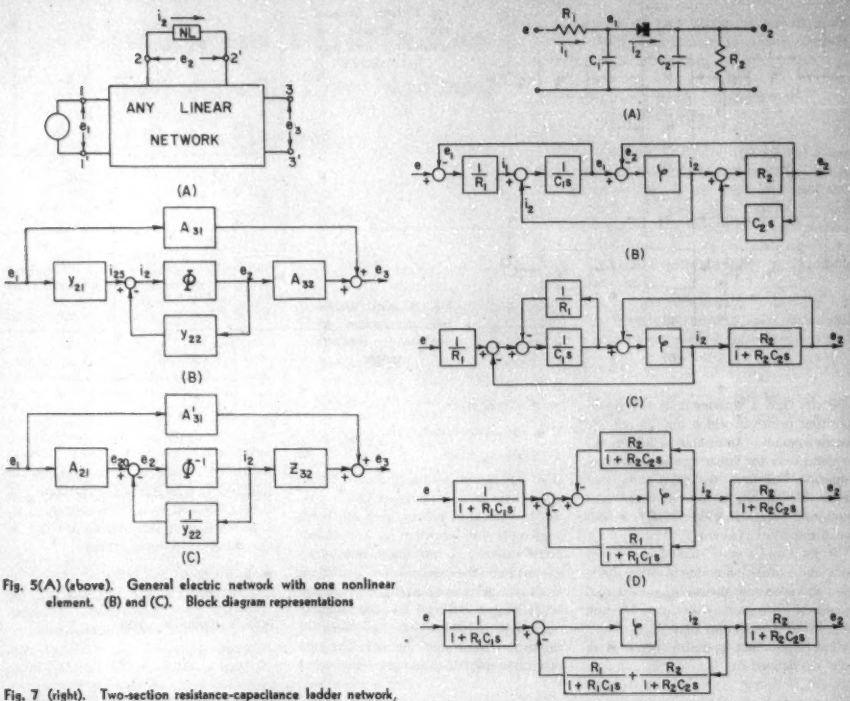


Fig. 5(A) (above). General electric network with one nonlinear element. (B) and (C). Block diagram representations

Fig. 7 (right). Two-section resistance-capacitance ladder network, with a diode rectifier and block diagrams

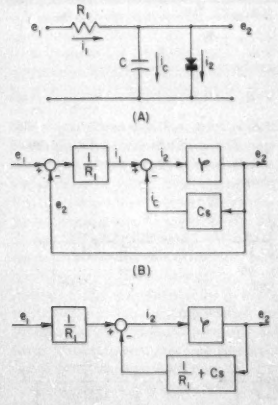


Fig. 6. One-section resistance-capacitance ladder network with a diode rectifier and block diagrams

Rule 7 describes a feedback system with a forward operator A_1 and unity feedback; the minus sign at the summing point indicates that the feedback is meant to be basically negative or degenerative. If A_1 represents a linear operator, the equivalent operator can be found by eliminating x from the equations

$$x = u - y \tag{11}$$

$$y = A_1 . x \tag{12}$$

If A_1 is a nonlinear operator, there are by definition no dynamic effects in the loop and an operator can be found to express the relation between the instantaneous values of y and u. Perhaps the simplest procedure is to assume a value of y, to find the corresponding x from a plot of $A_1(x)$, and then to find $u = x + y$. These steps can be carried out graphically in a systematic manner with the use of a special form of the more general procedure described in connection with rule 8. If A_1 is any mixture of linear operators with one nonlinear operator, the original diagram can be converted by use of rules 1 through 6 into the form considered in rule 8.

The original diagram of rule 8 describes a feedback system with a forward operator A_1 and a feedback operator A_2. If both operators are linear; the equivalent operator is found by eliminating x and v from the equations

$$x = u - v \tag{13}$$

$$y = A_1 . x \tag{14}$$

$$v = A_2 . y \tag{15}$$

If both operators are nonlinear (or if one denotes multiplication by a constant), an equivalent operator can be found in the following way. Note that

$$u = x + v \tag{16}$$

and determine x and v for a common value of y from

$$y = A_1(x) \tag{17}$$

$$v = A_2(y) \tag{18}$$

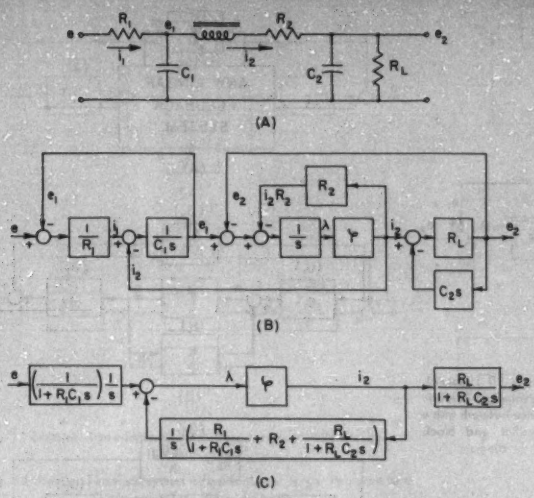


Fig. 8. Ladder network with a nonlinear inductor and block diagrams

Thereby u as a function of y is obtained, and conversely. To carry out this process graphically, the plot of $v=A_2(y)$ is first rotated counterclockwise through a 90-degree angle, as shown in Fig. 3(B). This plot is then superimposed on the plot of $y=A_1(x)$, shown separately in Fig. 3(A), to obtain the combined plot shown in Fig. 3(C). A horizontal line is then drawn through the intersection of the two curves. A point on the $y=A(u)$ curve is determined by the intersection of this horizontal line with the y axis of the v-y plot. The horizontal spacing of the two plots can be varied to change the location of the intersection of the two curves and, thereby, the value of y under consideration is changed; in this manner, the entire curve can be constructed. (This construction can be used to show the linearization and loss of output amplitude resulting from feedback and might have facilitated discussion of these matters by Rowlands,[10] and Zepler,[11] and others.)

If A_1 and A_2 represent a mixture of a nonlinear operator with linear operators, no equivalent operator can be defined and no further reduction of the diagram is possible. Methods for transient or frequency-response analysis can, however, be developed for systems described by Fig. 1, in which the feedback system of rule 8 presents the greatest difficulties.

Rule 9, shown in Fig. 4(A), is applicable to linear, nonlinear, or mixed systems and is justified by the fact that the system equations may be written in either of two ways

$$\begin{aligned} x &= u - v \\ y &= A_1 x \\ v &= A_2 y \end{aligned} \qquad (19)$$

or

$$\begin{aligned} v &= u - x \\ y &= A_2^{-1} v \\ x &= A_1^{-1} y \end{aligned} \qquad (20)$$

If, as shown in Fig. 4(B), A_1 is nonlinear and represents $y=\varphi(x)$, the inverse operator is the inverse function $x=\psi(y)$; the inverse of the linear operator $A_2=H$ is the reciprocal transfer function $1/H$. Rule 9 establishes the conditions for the equivalence of the two diagrams shown in Fig. 1 and insures that any system which can be represented by one of the standard diagrams can also be represented by the other.

Block Diagrams for Electric Networks

General Case

A general electric network containing a single nonlinear element is shown in Fig. 5(A). It is assumed that a voltage e_1 is applied at terminals 1 to 1′ and that the resulting voltage at terminals 3 to 3′ is to be found; the nonlinear element is connected between terminals 2 to 2′. (A similar development can be carried out if the excitation is a current source at terminals 1 to 1′.) The voltage-current relation for the nonlinear element is written as $e_3=\Phi(i_2)$, which implies that the element is a nonlinear resistor or diode; other types of nonlinear elements the voltage-current relations of which involve a differentiation or integration will require a minor modification of the pro-

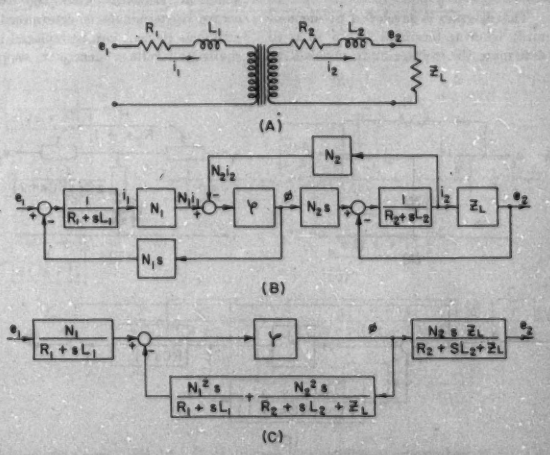
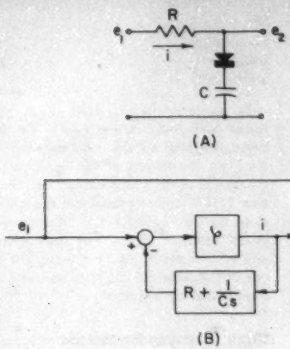
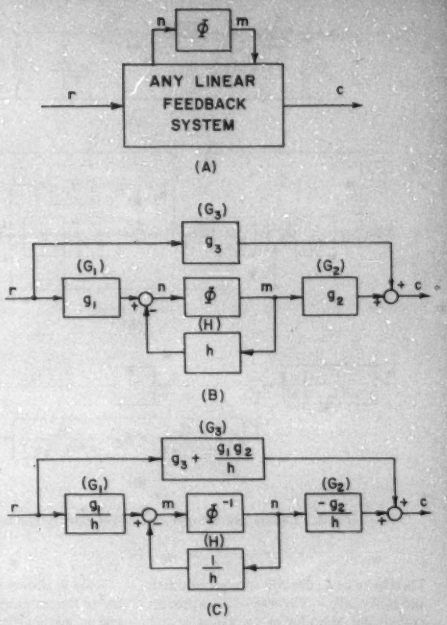


Fig. 9. Transformer with a nonlinear core material and block diagrams





cedure which alters the nature of the transfer functions but not the appearance of the final block diagram.

A possible block diagram for this network is shown in Fig. 5(B), in which the symbols have the following significance:

i_{2s} = the short-circuit current at terminals 2 to 2′ caused by e_1

$y_{21} = \frac{i_{2s}}{e_1}$ = the transfer admittance between terminals 2 to 2′ and 1 to 1′

y_{22} = the input admittance at terminals 2 to 2′ with terminals 1 to 1′ short-circuited

A_{11} = the ratio of the voltage at terminals 3 to 3′ to the applied voltage at terminals 1 to 1′ with terminals 2 to 2′ short-circuited

A_{12} = the ratio of the voltage at terminals 3 to 3′ to the voltage at terminals 2 to 2′ with terminals 1 to 1′ short-circuited

This diagram is developed by momentarily ignoring terminals 3 to 3′ so as to determine the voltage at terminals 2 to 2′. To do this, the short-circuit current i_{2s} is found and used as the input current to a parallel circuit composed of two branches: the input admittance of the network at terminals 2 to 2′ and the nonlinear element. After the voltage across its terminals is determined, the nonlinear element can be replaced by an equivalent voltage generator supplying the same voltage to the network. The system then becomes a linear network driven by voltage sources at terminals 1 to 1′ and 2 to 2′, and the superposition principle may be used to compute the desired output voltage at terminals 3 to 3′.

Fig. 10 (above). Resistance - capacitance network with a rectifier and block diagram

Fig. 12 (right). General feedback system with one nonlinear element and block diagram representations

The mathematical relations for this network can also be represented by the diagram of Fig. 5(C) in which:

e_{2o} = the open-circuit voltage at terminals 2 to 2′ caused by e_1

A_{21} = the ratio of the open-circuit voltage at terminals 2 to 2′ to the applied voltage at terminals 1 to 1′

$\frac{1}{y_{22}}$ = the input impedance at terminals 2 to 2′ with terminals 1 to 1′ short-circuited

A_{21}' = the ratio of the voltage at terminals 3 to 3′ to the applied voltage at terminals 1 to 1′ with terminals 2 to 2′ open-circuited

z_{23} = the ratio of the voltage at terminals 3 to 3′ to the current i_2 with terminals 1 to 1′ short-circuited

This diagram is developed by first finding the current in a series circuit consisting of the nonlinear element and the input impedance at terminals 2 to 2′, excited by the open-circuit voltage appearing there. The nonlinear element is then replaced by an equivalent current generator. The output voltage at terminals 3 to 3′ is computed as a superposition of

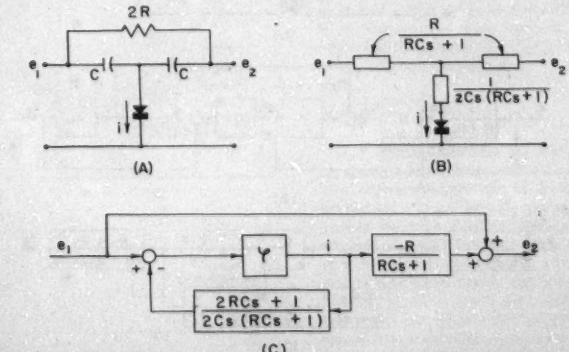


Fig. 11(A). Bridged-T network with a rectifier. (B). Equivalent circuit. (C). Block diagram

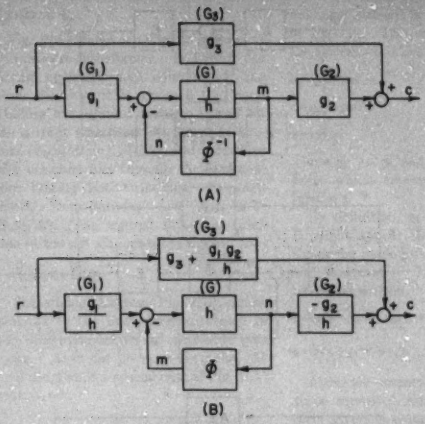


Fig. 13 (above). Two additional block diagram representations of general feedback system

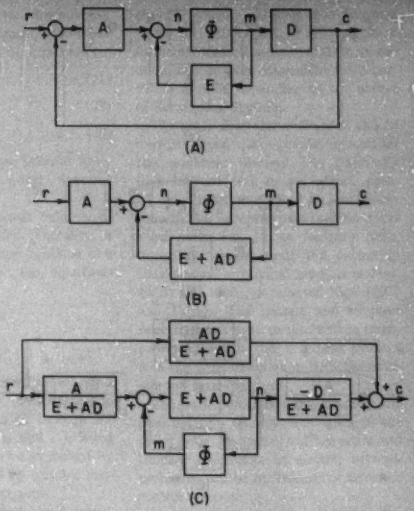


Fig. 15 (right). Feedback system with a nonlinear forward element and two equivalent block diagrams

components caused by a voltage source at terminals 1 to 1′ and a current source at terminals 2 to 2′.

The diagrams of Fig. 5(B) and 5(C) fit the general form of Fig. 1(A). For Fig. 5(B), i_2 and e_2 are identified with x and y respectively, while the identification is reversed for Fig. 5(C); the significance of the linear blocks is different in the two diagrams. Two more representations of the network of Fig. 5(A) can be obtained by use of rule 9, giving diagrams which fit the form of Fig. 1(B); in these transformations, the significance of the blocks G_1, G_2, and G_3 is maintained. Which block diagram provides the simplest and neatest representation of a particular network depends on the nature of the network, just as mesh or nodal equations may be more convenient for analysis of a particular linear circuit.

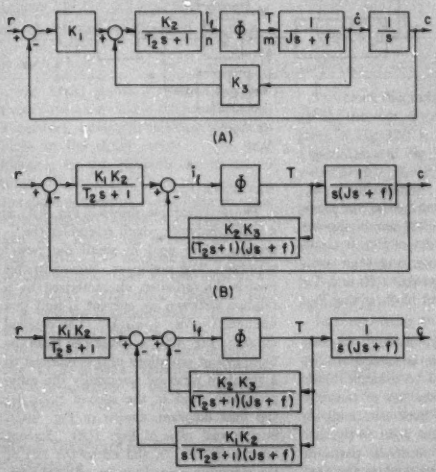


Fig. 14. Block diagrams for a positioning servomechanism with a nonlinear torque-current relation

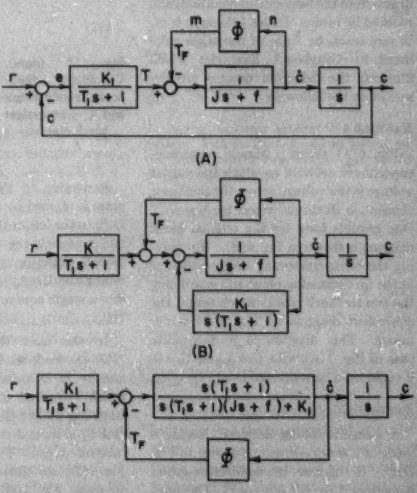
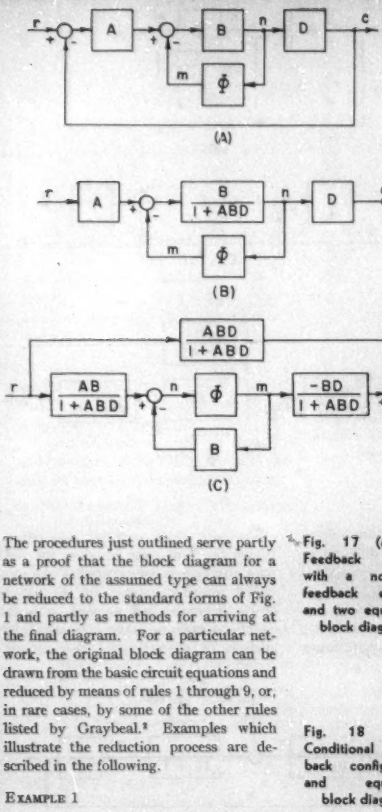


Fig. 16. Block diagrams for a positioning servomechanism with a nonlinear friction torque



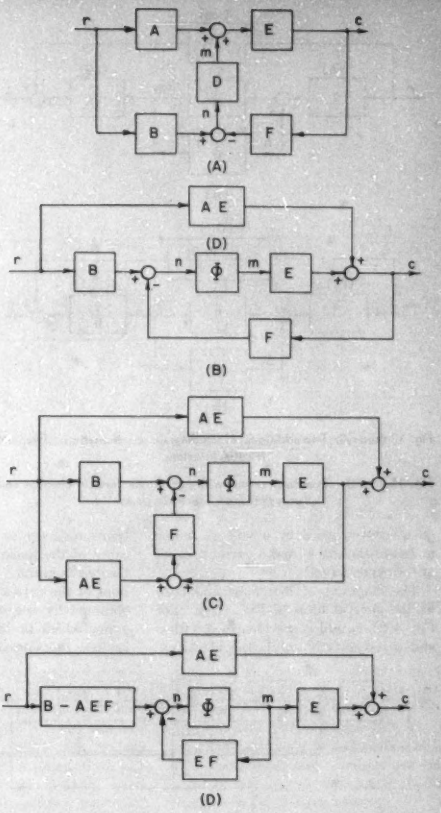


Fig. 17 (above). Feedback system with a nonlinear feedback element and two equivalent block diagrams

Fig. 18 (right). Conditional feedback configuration and equivalent block diagrams

The procedures just outlined serve partly as a proof that the block diagram for a network of the assumed type can always be reduced to the standard forms of Fig. 1 and partly as methods for arriving at the final diagram. For a particular network, the original block diagram can be drawn from the basic circuit equations and reduced by means of rules 1 through 9, or, in rare cases, by some of the other rules listed by Graybeal.[2] Examples which illustrate the reduction process are described in the following.

EXAMPLE 1

Fig. 6(A) shows a simple resistance-capacitance network in which the output voltage is the voltage across the nonlinear element, a diode described by $e_2=\varphi(i_2)$. One possible form for the original block diagram is shown in Fig. 6(B). By moving the block containing $1/R_1$ to the left of the first summing point and combining the two feedback blocks which result, the equivalent diagram of Fig. 6(C) is obtained. This diagram is a degenerate case of Fig. 1(A) with $G_3=1$ and $G_2=0$; it could have been derived from Fig. 5(B).

EXAMPLE 2

A 2-section ladder network, having a diode as a series element, is shown in Fig. 7(A). In this case, the nonlinear relation is expressed by $i_2=\varphi(e_1-e_2)$. The original block diagram, Fig. 7(B), is developed directly from the circuit equations.

Reduction of the original block diagram is started by first moving the block $1/R_1$ to the left of the first summing point and reversing the order of the first and second summing points and then combining the blocks containing R_2 and C_2s into a single equivalent block, giving Fig. 7(C).

In the next step, leading to Fig. 7(D), the feedback loop consisting of $1/R_1$ and $1/C_1s$ is reduced to a single block which is shifted to the left of the first summing point; the block containing R_2 and C_2 is moved to the right of the last take-off point. The feedback paths in Fig. 7(D) are added to produce the final diagram, Fig. 7(E). This diagram is a special case of Fig. 1(A), with $G_3=0$, and could have been obtained from Fig. 5(C).

EXAMPLE 3

The next example, shown in Fig. 8(A), is a ladder network which resembles the π-section filters used in small d-c power supplies. The nonlinear element in this case is the inductor, characterized by a relation between its current i_2 and flux linkages $\lambda=N\phi$; $i_2=\varphi(\lambda)$. The original diagram, Fig. 8(B), is developed from the basic circuit equations and is reduced to a standard form by essentially the same procedure used in the second example; the final diagram, shown in Fig. 8(C), is a special case of Fig. 1(A). Except for the factor $1/s$, shifted to the left of the summing point from its original position preceding the nonlinear block φ, the blocks represent voltage ratios and impedances; the final diagram could have been obtained by use of Fig. 5(C).

EXAMPLE 4

Fig. 9(A) shows a network containing an iron-core transformer. In writing the circuit equations and constructing the block diagram, it is assumed that each winding has induced voltages caused by the current variations in a linear inductance (L_1di_1/dt or L_2di_2/dt), and variations of a common flux through the nonlinear core linking both windings, computed directly from Faraday's law ($N_1d\phi/dt$ or $N_2d\phi/dt$). The mutual flux is a function of the net ampere-turns written as

$$\phi=\varphi(N_1i_1-N_2i_2)$$

The original diagram, Fig. 9(B), can be reduced to the final diagram shown in Fig. 9(C); the details are omitted to save space. As in the previous examples, the final diagram is a special case of Fig. 1(A) with $G_3=0$.

EXAMPLE 5

Fig. 10(A) shows a simple resistance-capacitance network containing a nonlinear diode. The final diagram for this network, Fig. 10(B), can be drawn directly by observing that e_2 can be computed from

$$e_2=e_1-iR$$

if i is known; i is simply the current in a series circuit composed of the diode and the linear elements R and C. The final diagram is a special case of Fig. 1(A), with $G_1=G_2=1$ and $G_3=-R$, and could have been obtained by use of Fig. 5(C).

EXAMPLE 6

Fig. 11(A) shows a resistance-capacitance network similar to the usual bridged-T network, having in this case a diode in place of the shunt resistor. The final diagram for this network can be drawn directly from Fig. 5(C) or by converting the delta connection of the linear resistor $2R$ and the two capacitors C to an equivalent wye, giving the network shown in Fig. 11(B). This network can be treated in the same manner as the network of Fig. 10(A). The final diagram, Fig. 11(C) fits the standard form of Fig. 1(A), with $G_1=G_2=1$, and $G_3=-R/(RCs+1)$. Even though the linear shunt element in Fig. 11(B) is not a realizable impedance, the representation is valid for analytical purposes.

Block Diagrams for Feedback Systems

GENERAL CASE

A general feedback system with a single input r, a single output c, and one nonlinear element is shown schematically in Fig. 12. The relationships between the system variables can be expressed by

$$n=g_1r-hm \tag{21}$$

$$c=g_2r+g_3m \tag{22}$$

$$m=\Phi(n) \tag{23}$$

where g_1, g_2, g_3, and h are transfer functions which represent linear dynamic relations.

By definition, m depends only on n. It might appear that, for generality, n should be made an explicit function of c and conversely. If so, the equations would be written

$$n=g_1'r-h_1'm-h_2'c$$

$$c=g_3'r+g_3'm+g_4'n$$

Algebraic manipulation shows that these equations can be reduced to the form given in equations 21 through 23. Any relations between n and c existing physically can therefore be expressed by equations 21 through 23 by suitable definition of the transfer functions.

Equations 21 through 23 lead immediately to the block diagram of Fig. 12(B), a special case of Fig. 1(A), with $x=n$, $y=m$, $\varphi=\Phi$, $G_1=g_1$, $G_2=g_2$, $G_3=g_3$, and $H=h$. Equations 21 through 23 can be rearranged to give

$$m=\frac{g_1}{h}r-\frac{1}{h}n \tag{21A}$$

$$c=\left(g_2+\frac{g_1g_3}{h}\right)r-\frac{g_3}{h}n \tag{22A}$$

$$n=\Phi^{-1}(m) \tag{23A}$$

The block diagram based on equations 21(A) through 23(A) takes the form given in Fig. 12(C); this diagram is also a special case of Fig. 1(A), but now $x=m$, $y=n$, $\varphi=\Phi^{-1}$, $G_1=g_1/h$, $G_2=-g_3/h$, $G_3=g_2+g_1g_3/h$, and $H=1/h$. These two diagrams are somewhat analogous to the two possible representations of the electric network given in Fig. 5(B) and 5(C). As before, two additional diagrams, shown in Fig. 13(A) and 13(B), can be derived by use of rule 9.

In constructing the original block diagram for an electric network, the individual elements can generally be characterized by their impedances or admittances with equal ease; it is immaterial whether the current is thought to produce the voltage or vice versa. The analyst does not feel the same flexibility in drawing the original block diagram for a feedback control system, however, since he usually tries to construct a diagram having a strong topological resemblance to the parts and interconnections in the physical system. Although the diagrams shown in Figs. 12 and 13 are equivalent in a mathematical sense, one form or another may be more convenient for a particular system, depending on the nature of the original diagram.

The block diagrams of Figs. 12 and 13 were developed without specifying whether the nonlinear element was physically in a forward path or a feedback path. The diagrams show that the nonlinear element can be represented in either a forward path, Fig. 12, or feedback path, Fig. 13, with its input and output in their normal or original sense as in Figs. 12(B) and 13(B), or reversed Figs. 12(C) and 13(A). The neatest and simplest final diagrams seem to result when transmission through the nonlinear element is in the same direction in both the original and final diagrams.

Short-cut procedures for obtaining the final diagram directly, analogous to the direct determination of voltage ratios and transfer emittances in electric network problems, can be developed for feedback control systems:

1. If the nonlinear element is a forward element, with input n and output m, the transfer functions of Fig. 12(B) can be evaluated from

$$\left.\frac{n}{m'}\right|_{r=0}=-h=-H$$

$$\left.\frac{n}{r}\right|_{m'=0}=g_1=G_1$$

$$\left.\frac{c}{m'}\right|_{r=0}=g_3=G_2 \tag{24}$$

$$\left.\frac{c}{r}\right|_{m'=0}=g_2=G_3$$

2. If the nonlinear element is a feedback element, with an input n and an output m, the linear transfer functions of Fig. 13(B) can be evaluated from

$$\left.\frac{n}{m'}\right|_{r=0}=-h=-G$$

$$\left.\frac{n}{r}\right|_{m'=0}=g_1=G_1G$$

$$\left.\frac{c}{m}\right|_{r=0}=-\frac{g_3}{h}=G_2 \tag{25}$$

$$\left.\frac{c}{r}\right|_{n=0}=g_2+\frac{g_1g_3}{h}=G_3$$

In carrying out these calculations, it is imagined that the input and output paths of the nonlinear block are opened and that a signal m' is supplied to the system in place of m. Note that $m=0$ does not necessarily mean $n=0$, and vice versa. The transfer function $-H$, for example, permits the calculation of n for a specified

m (with $r=0$) because of transmission through only the linear part of the system. These transfer functions are *not* the same as the loop gains defined by Mason[8] or the return differences of Truxal[9] and Bode,[13] since the latter refer to transmission completely around a loop back to the starting point, with none of the elements being ignored. Examples illustrating the use of the transformation rules and the short-cut procedures follow.

EXAMPLE 7

Fig. 14(A) shows the original block diagram of a tachometer-stabilized positioning servomechanism in which the d-c motor torque T is assumed to be a nonlinear function of the control field current i_f. The nonlinearity is represented by a block in a forward path.

Reduction of this diagram is started by moving $K_2/(T_2s+1)$ to the left of the second summing point and $1/(Js+f)$ to the right of the first take-off point, which gives the diagram of Fig. 14(B). The block $K_1K_2/(T_2s+1)$ is then moved to the left of the first summing point and $1/s(Js+f)$ is moved to the right of the last take-off point, which gives Fig. 14(C). Combination of the two feedback blocks gives a final diagram (not shown) which is a special case of Fig. 1(A), with $G_3=0$.

EXAMPLE 8

Fig. 15(A) shows a general block diagram having the same form as Fig. 14(B). Direct application of equation 24 to develop a diagram similar to Fig. 12(B) gives

$$\left.\frac{n}{m'}\right|_{r=0} = -h = -H = -(E+AD)$$

$$\left.\frac{n}{r}\right|_{m'=0} = g_1 = G_1 = A$$

$$\left.\frac{c}{m'}\right|_{r=0} = g_2 = G_2 = D \tag{26}$$

$$\left.\frac{c}{r}\right|_{m'=0} = g_3 = G_3 = 0$$

resulting in the diagram of Fig. 15(B). On the other hand, equation 25 gives

$$\left.\frac{n}{m'}\right|_{r=0} = -G = -(E+AD)$$

$$\left.\frac{n}{r}\right|_{m'=0} = G_1G = A$$

$$G_1 = \frac{A}{E+AD} \tag{27}$$

$$\left.\frac{c}{n}\right|_{r=0} = \frac{\left.\frac{c}{m'}\right|_{r=0}}{\left.\frac{n}{m'}\right|_{r=0}} = G_2 = \frac{-D}{E+AD}$$

$$\left.\frac{c}{r}\right|_{n=0} = G_3 = \frac{AD}{E+AD}$$

the last transfer function being obtained from the relation

$$n = A(r-c) - Em' = A(r-c) - \frac{E}{D}c = 0$$

The same results can be obtained by substituting g_1, g_2, g_3, and h from equation 26 into equation 25; the diagram based on equation 27 is shown in Fig. 15(C). Even without attempting to substitute specific transfer functions for A, D, and E, it is evident that the simplest representation of this system is given by Fig. 15(B).

EXAMPLE 9

Fig. 16(A) shows the original block diagram of a positioning servomechanism having a friction torque T_F which is a nonlinear function of the output velocity $\dot{c}$.

Reduction of the block diagram is started by moving $K_1/(T_1s+1)$ to the left of the first summing point, changing the order of the summing points, and moving $1/s$ to the right of the last take-off point, which gives Fig. 16(B). The feedback loop at the center of the diagram is then reduced to an equivalent single block, which gives Fig. 16(C), a special case of Fig. 1(B).

EXAMPLE 10

Fig. 17(A) shows a general block diagram having the same form as Fig. 16(A). Direct application of equation 25 to develop a diagram similar to Fig. 13(B) gives

$$\left.\frac{n}{m'}\right|_{r=0} = -h = -G = \frac{-B}{1+ABD}$$

$$\left.\frac{n}{r}\right|_{m'=0} = g_1 = G_1G = \frac{AB}{1+ABD}$$

$$G_1 = A \tag{28}$$

$$\left.\frac{c}{n}\right|_{r=0} = -\frac{g_2}{h} = G_2 = D$$

$$\left.\frac{c}{r}\right|_{n=0} = g_3 + \frac{g_1g_2}{h} = G_3 = 0$$

which results in the diagram of Fig. 17(B).

Equation 28 can be used to obtain

$$h = \frac{B}{1+ABD}$$

$$g_1 = \frac{AB}{1+ABD}$$

$$g_2 = \frac{-BD}{1+ABD} \tag{29}$$

$$g_3 = \frac{ABD}{1+ABD}$$

for substitution into Fig. 12(B), which gives the diagram shown in Fig. 17(C). The simplest representation of this system is given by Fig. 17(B).

EXAMPLE 11

The final example is based on the "conditional feedback" configuration suggested by Lang and Ham[13] and is shown, with some changes in notation, in Fig. 18(A).

The original diagram can be reduced to a standard form very easily when either A, B, E, or F is made a nonlinear block. If D becomes nonlinear, say $m=\Phi(n)$, the process is more difficult. Reduction of the diagram can be started by moving E to the left of the upper summing point and redrawing the diagram in the form given in Fig. 18(B). The next step, resulting in Fig. 18(C), is the interchange of the final summing and take-off points; this requires the introduction of a new summing point so that the input to F remains $(AEr+Em)$. Shifting F past the lower summing point and changing the order of the first two summing points results in two parallel forward blocks B and $-AEF$, which can be combined into a single block, and a feedback path around Φ and E, which can be reduced to the standard form by shifting E past the right-hand take-off point. The final diagram is shown in Fig. 18(D).

Fig. 18(A) can be reduced directly to the form shown in Fig. 12(B) by using equation 24, since

$$\left.\frac{n}{m'}\right|_{r=0} = -H = -EF$$

$$\left.\frac{n}{r}\right|_{m'=0} = G_1 = B - AEF \tag{30}$$

$$\left.\frac{c}{m'}\right|_{r=0} = G_2 = E$$

$$\left.\frac{c}{r}\right|_{m'=0} = G_3 = AE$$

A diagram (not shown) having the form of Fig. 13(B) can be drawn with the following:

$$G = EF$$

$$G_1 = \frac{B-AEF}{EF} \tag{31}$$

$$G_2 = -\frac{1}{F}$$

$$G_3 = \frac{B}{F}$$

Since D may be regarded as either a forward or feedback block in this case, the two diagrams are about equally complex.

Conclusions

The usual block diagram transformation rules can be applied to systems containing nonlinear elements, provided that certain restrictions are observed. Nonlinear blocks cannot be moved past summing points; the order of linear and nonlinear blocks cannot be reversed; and feedback loops cannot be reduced to equivalent blocks (except in special cases). Variables can be combined by addition and subtraction in any required order, and linear blocks obey the same rules as before, so that considerable modification and reduction of the block diagram is still possible.

Systems with one input, one output, and a nonlinear element characterized by $y=\varphi(x)$ can always be represented by four block diagrams which are topologically identical and mathematically equivalent. In these diagrams, which may be obtained by use of the transformation rules or by more direct procedures, the nonlinear element may appear as either a forward or feedback element with its input and output in either a normal or reversed cause-effect relationship.

Block diagrams transformations serve as a substitute for algebraic manipulations of the system equations used, e.g., by Ku in his work on higher-order nonlinear systems.[14] The block diagram approach has the advantage of maintaining close contact with the physical system, providing new insights into system operation, and suggesting formulations of nonlinear problems which may facilitate further analysis. The last point is particularly important since some of the difficulties in analyzing nonlinear systems may result from a poor choice of variables or from use, by force of habit, of an inconvenient description of the nonlinearity.

References

1. Principles of Servomechanisms (book), G. S. Brown, D. P. Campbell. John Wiley & Sons, Inc., New York, N. Y., 1948.

2. Block Diagram Network Transformation, T. D. Graybeal. *Electrical Engineering*, vol. 70, Nov. 1951, pp. 985–90.

3. A Block Diagram Approach to Network Analysis, T. M. Stout. AIEE *Transactions*, vol. 71, pt. II, 1952, pp. 255–60.

4. Block Diagram Solutions for Vacuum-Tube Circuits, T. M. Stout. *Ibid.*, vol. 72, pt. I, Nov. 1953, pp. 561–67.

5. Signal Flow Diagrams for Process Evaluation, D. P. Campbell. *Industrial and Engineering Chemistry*, American Chemical Society, Easton, Pa., vol. 47, Mar. 1955, pp. 409–13.

6. A Frequency Response Method for Analyzing and Synthesizing Contactor Servomechanisms, Ralph J. Kochenburger. *AIEE Transactions*, vol. 69, pt. I, 1950, pp. 270–84.

7. Describing Function Method of Servomechanism Analysis Applied to Most Commonly Encountered Nonlinearities, H. D. Greif. *Ibid.*, vol. 72, pt. II, Sept. 1953, pp. 243–48.

8. Feedback Theory—Some Properties of Signal Flow Graphs, S. J. Mason. *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 41, Sept. 1953, pp. 1144–56.

9. Automatic Feedback Control System Synthesis (book), J. G. Truxal. McGraw-Hill Book Company, Inc., New York, N. Y., 1955, pp. 88–160.

10. Harmonic Distortion and Negative Feedback, R. O. Rowlands. *Wireless Engineer*, London, England, vol. 30, June, 1953, pp. 133–35. Discussions: Sept., pp. 232–33; Oct., pp. 261–62; Nov., pp. 291–92.

11. Harmonic Distortion and Negative Feedback, E. E. Zepler. *Ibid.*, vol. 31, May, 1954, pp. 118–21. Discussion: Nov. 1954, p. 306.

12. Network Analysis and Feedback Amplifier Design (book), H. W. Bode. D. Van Nostrand Company, Inc., New York, N. Y., 1945.

13. Conditional Feedback Systems—A New Approach to Feedback Control, G. Lang, J. M. Ham. *AIEE Transactions*, vol. 74, pt. II, Jul. 1955, pp. 152–61.

14. Analysis of Nonlinear Coupled Circuits, Y. H. Ku. *Ibid.*, vol. 73, pt. I, Jan. 1955, pp. 626–31.

Discussion

Abraham M. Fuchs (Westinghouse Electric Corporation, Baltimore, Md.): In a previous article[1] a further transformation of the block diagram of a system containing one nonlinear element is carried out. The control loop containing the nonlinearity is replaced by two blocks in series. One block is the closed loop equivalent linear system, while the second is a new nonlinearity which operates to produce the same output-input of the two blocks in series as the original closed loop system.

The basis for this transformation was the assumption that the describing function of the nonlinearity was an adequate representation of the nonlinearity. The results were limited to nonlinearities whose describing functions have no phase shift. The equivalent series nonlinearity will have phase shift and will also be a function of frequency.

I feel that this extension of the block diagram transformation for many systems with one nonlinear element is a valuable tool in feedback control system synthesis. It provides a straightforward approach to analyzing the outer loop stability when the original nonlinearity is part of an inner loop. Additions to this procedure for a more general class of nonlinearities may be possible.

Reference

1. Non-Linear Boost System Flow Characteristics, A. M. Fuchs, F. J. Huddleston. *Aeronautical Electronics Digest*, Institute of Radio Engineers, New York, N. Y., pp. 242–50.

L. A. Zadeh (Columbia University, New York, N. Y.): The purpose of this discussion is to outline a method of approach which leads to the standard form obtained by Dr. Stout somewhat more easily and which has the advantage of being readily applicable to systems containing more complex nonlinear elements.

Specifically, any 2-pole (i.e., device having one input and one output) which contains a nonlinear 2-pole can be regarded as a combination of a linear (2,2)-pole (i.e., device with two inputs and two outputs) and a nonlinear 2-pole; see Fig. 19. It should be noted that the nonlinear 2-pole need not be of the zero-memory type, as assumed by Dr. Stout.

Now, any linear (2,2)-pole can be represented as a combination of four linear 2-poles and 2-adders, as shown in Fig. 20. On replacing the linear (2,2)-pole in Fig. 19 by that shown in Fig. 20, one obtains the system shown in Fig. 21. This system is identical with the standard form obtained by the author, the only difference being the adder P which in the author's form is a subtractor. This difference, however, is superficial, since any adder may be replaced by a subtractor and vice versa, merely by multiplying the transmission functions of appropriate blocks by minus one.

This simple approach can readily be used in the case of systems containing more complex elements. For example, if a 2-pole system contains a single nonlinear (m,n)-pole (i.e., a device with m inputs and n outputs) then it may be regarded as a combination of a linear $(n+1, m+1)$-pole and a nonlinear (m,n)-pole. Now a linear $(n+1, m+1)$-pole may be decomposed into $(m+1)(n+1)$ linear 2-poles and $m+1$ adders. Furthermore, any nonlinear (m,n)-pole can be decomposed into n nonlinear $(m,1)$-poles.[1] Consequently, any 2-pole containing a single nonlinear (m,n)-pole can be represented as a combination of n nonlinear $(m,1)$-poles, $(m+1)(n+1)$ linear 2-poles and $n+1$ adders (or subtractors).

Reference

1. Nonlinear Multipoles, L. A. Zadeh. *Proceedings*, National Academy of Sciences, Washington, D. C., vol. 39, April 1953, pp. 274–80.

T. M. Stout: The discussers have made useful additions to the paper. To clarify the block diagram transformation discussed by Mr. Fuchs, a diagram from reference 1 of his discussion is reproduced in Fig. 22. Since they are concerned only with sinusoidal inputs, the nonlinear element is characterized by a describing function $N(\alpha)$ and the linear elements by a function $G(\omega)$ giving the magnitude ratio and phase shift as functions of frequency. The closed loop gains

$$G_1(\alpha,\omega)=\frac{C}{R}=\frac{G(\omega)}{\dfrac{1}{N(\alpha)}+G(\omega)} \tag{32}$$

$$G_2(\omega)=\frac{C}{R}=\frac{G(\omega)}{1+G(\omega)} \tag{33}$$

are computed. After these computations, the equivalent series nonlinearity

$$N'(\alpha,\omega)=\frac{G_1(\alpha,\omega)}{G_2(\omega)} \tag{34}$$

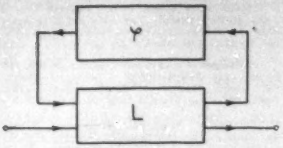


Fig. 19. Combination of linear (2,2-pole) and nonlinear 2-pole

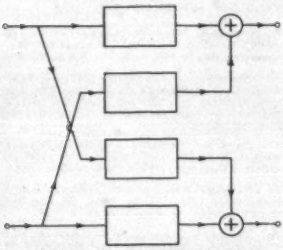


Fig. 20. Linear 2,2-pole represented as combination of four linear 2-poles and 2-adders

is determined. Since, by definition

$$G_1(\alpha,\omega) = G_2(\omega)N'(\alpha,\omega) \qquad (35)$$

the systems of Fig. 22(A) and (B) are equivalent.

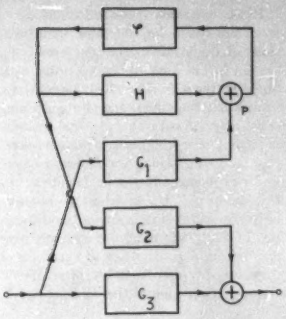


Fig. 21. System obtained by replacing linear (2,2-pole) of Fig. 19 by that in Fig. 20

Representation of the complete system by the diagram of Fig. 22(B) may have conceptual value, but its computational utility can be questioned. If this system is part of a larger system and further computations are required, it would appear more convenient to use $G_1(\alpha,\omega)$, defined by equation 32, than $N'(\alpha,\omega)$. As indicated, both are functions of amplitude and frequency, and use of G_1 eliminates the computations indicated in equations 33 and 34.

It should be emphasized that the transformation described by Mr. Fuchs is applicable only for sinusoidal inputs and that, even in this case, the calculated C/R ratio is an approximation. In discussing the application of rules for block diagram

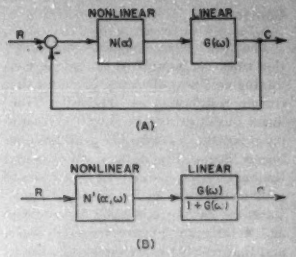
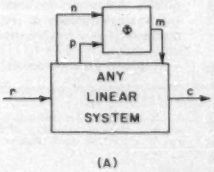
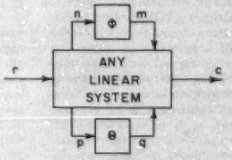


Fig. 22. Block diagram transformation from reference 1 of Mr. Fuch's discussion

transformation to nonlinear systems, the criterion employed in this paper is their validity for all kinds of input signals. This criterion clearly restricts the amount of diagram modification which is permissible. No attempt has been made here to determine which rules are valid for special inputs or what transformations can be made by permitting the new system to approximate the original in some manner. The question of approximate representation is explored in the paper mentioned by Prof. Zadeh, one of a number of unique and stimulating contributions he has made to this field.

The approach suggested by Prof. Zadeh provides an interesting alternative to the method given in the paper. It does, however, seem to require prior proof of the separate statements that "any 2-pole which contains a nonlinear 2-pole can be regarded as a combination of a linear (2,2)-pole and a





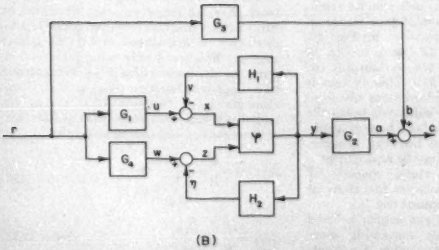


Fig. 23 (above). A—System containing nonlinear relation $m = \Phi(n,p)$. B—Corresponding general diagram

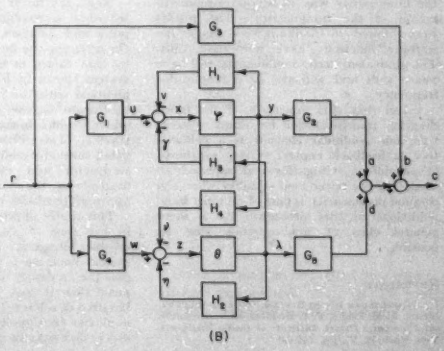


Fig. 24 (right). A—System containing two nonlinear relations $m = \Phi(n)$ and $q = \Theta(p)$. B—Corresponding general diagram

nonlinear 2-pole," and "any linear (2,2)-pole can be represented by a combination of four linear 2-poles and 2-adders." Although these statements are not hard to prove, it seems equally direct to write the relations for the complete system and to draw the corresponding diagram.

The method employed in this paper can also be applied to more complex systems, although this had not been done at the time the paper was written. A system containing a nonlinear relation $m=\Phi(n,p)$, shown in Fig. 23(A), can be represented by the general diagram of Fig. 23(B). The identity of these diagrams is established by writing the relations for the original system in the following manner:

$$\begin{aligned} n &= g_1 r - h_1 m \\ p &= g_2 r - h_2 m \\ c &= g_3 m + g_3 r \\ m &= \Phi(n,p) \end{aligned} \tag{36}$$

Setting $n=x$, $p=z$, $m=y$, and $\Phi=\varphi$, and writing G for g and H for h, these equations lead directly to Fig. 23(B). The similarity of this figure to Fig. 12 is evident. The dual notation, which may seem unnecessary, is employed deliberately here and in the paper. It should serve as a reminder that the final diagram is not unique and that, as shown in the paper for the simpler systems, several alternative representations of the same system are available.

A system containing two nonlinear relations, $m=\Phi(n)$ and $q=\Theta(p)$, is shown in Fig. 23(A). The corresponding general diagram, developed in the same manner as before, is given in Fig. 24(B). Similar procedures can be applied to more complicated systems, but the diagrams grow rapidly in size and complexity.

In Prof. Zadeh's terminology, the system of Fig. 23(A) is a 2-pole containing a nonlinear (2,1)-pole. His concluding paragraph indicates that this system could be represented by a linear (3,2)-pole and a nonlinear (2,1)-pole, or by (3)(2)=6 linear 2-poles, 3 adders, and a nonlinear (2,1)-pole, exactly as shown in Fig. 23(B). Likewise, the system shown in Fig. 24(A) contains two nonlinear 2-poles or, what amounts to the same thing, a degenerate nonlinear (2,2)-pole. Its representation requires (3)(3)=9 linear 2-poles and 3 adders. Since $c=a+b+d$ actually requires only one adder, the diagram shown in Fig. 24(B) agrees with his statements. These examples may help to explain the significance and application of the general results obtained by Prof. Zadeh.

The term "pole" is somewhat unfortunate, since the Institute of Radio Engineers definitions of terms[1] defines a 4-pole as a 2-terminal pair network (usually a network with one input and one output). The term "port," already suggested as a substitute for "terminal pair," might be preferable.

Use of adders or subtractors is indeed superficial. In Figs. 10, 11, 15, and 17, negative signs are introduced for reasons suggested by Prof. Zadeh. In the general diagrams, subtractors are used in feedback loops where the feedback is generally expected to be negative, in order to avoid associating negative signs with the H's.

Restriction to zero-memory nonlinearities is not essential for the block diagram representation or transformation. It did seem necessary for a method for computing transient response which is based on the standard diagram, described in another paper.[2]

References

1. Network Topology. *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 39, Jan. 1951, pp. 27–29.

2. A Step-by-Step Method for Transient Analysis of Feedback Systems with One Nonlinear Element, T. M. Stout. *AIEE Transactions* vol. 75, pt. II, 1956 (*Paper no. 56-779*).

Synthesis and Critical Study of Sampled-Data Control Systems

ELIAHU I. JURY
ASSOCIATE MEMBER AIEE

Synopsis: In the last few years, there have been few papers[1-9] which discuss methods of analysis and synthesis of sampled-data control systems. In this paper it is shown that the z-transform method is the best as applied to the design of such systems. The z-transform method yields information only on the response of the system at sampling instants. Certain methods have been introduced which modify the z-transform method to obtain information on the system's behavior between sampling instants. All these proposed methods are mainly directed for analysis purposes. However, none is applied for design procedures which can be consolidated with the z-transform method. In this work, the modified[1] z-transform method is further extended and applied for synthesis procedures. It is believed that the method chosen is best applicable for the design of sampled-data control systems. Both the frequency-response and the root-locus approach with the use of this modified z-transform method are discussed. Furthermore, the effect of pole and zero location on the transient response of sampled-data systems is investigated, and certain information on the response obtained.

The correlation is obtained between frequency and transient overshoot, and the results tabulated for different parameters of the system. Constant overshoot loci, which can be used for design and synthesis procedures, are plotted in the z-plane. A few examples are introduced, illustrating and applying the theoretical results.

THE conclusion is reached that, with the advent of the modified z-transform method, the limitations of the z-transform method are completely removed and that it is now possible to design any sampled-data control system to obtain all of the design information to complete analogy with continuous systems.

Mathematical Background

It is known that the z-transform method yields the response of a sampled-data system at sampling instants only. However, to obtain the response between sampling instants, one may introduce a fictitious delay Δ in the forward path, e.g., when Δ varies from zero to unity, the total response between sampling instants may be obtained.

The z-transform of the sampled-data system transfer function in this case is given as

$$G^*(z,\Delta)=\mathcal{L}[g(t-\Delta T)\times\delta_T(t)] \tag{1}$$

Effecting the contour integration in the left half of the plane, as shown in Fig. 3 of reference 5, one may obtain the $G^*(s,\Delta)$ function, as follows

$$G^*(z,\Delta)=\sum_{\substack{\text{poles}\\ \text{of } G(s)}} \text{residue of } \left(G(s)\frac{e^{-\Delta sT}}{1-e^{+sT}z^{-1}}\right) \tag{2}$$

where

$$G(s)e^{-\Delta sT}=\mathcal{L}[g(t-\Delta T)] \tag{3}$$

If the notation $m=1-\Delta$ is introduced, then the modified pulsed transfer function becomes

$$G^*(z,m)=z^{-1}\sum_{\substack{\text{poles}\\ \text{of } G(s)}} \text{residue of } \left(G(s)\frac{e^{msT}}{1-e^{+sT}z^{-1}}\right) \tag{4}$$

Paper 56-208, recommended by the AIEE Feedback Control Systems Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Winter General Meeting, New York, N. Y., January 30-February 3, 1956. Manuscript submitted August 18, 1955; made available for printing November 23, 1955.

Eliahu I. Jury is with the University of California, Berkeley, Calif.

Table I. Pairs of Modified z-Transforms*

$G(s)$	$G^*(z, m)$
$\frac{1}{s}$	$\frac{1}{z-1}$, $\Delta>0$
$\frac{1}{s^2}$	$\frac{mT}{z-1}+\frac{T}{(z-1)^2}$, $m=(1-\Delta)$
$\frac{1}{s^3}$	$\frac{T^2}{2}\left(\frac{m^2}{z-1}+\frac{2m+1}{(z-1)^2}+\frac{2}{(z-1)^3}\right)$
$\frac{1}{s+a}$	$\frac{e^{-amT}}{z-e^{-aT}}$
$\frac{1}{(s+a)^2}$	$Te^{-amT}\frac{m}{z-e^{-aT}}+\frac{e^{-aT}}{(z-e^{-aT})^2}$
$\frac{1}{(s+a)^3}$	$\frac{T^2e^{-mTa}}{2}\frac{m^2}{z-e^{-aT}}+\frac{(2m+1)e^{-aT}}{(z-e^{-aT})^2}+\frac{2e^{-aT}}{(z-e^{-aT})^3}$
$\frac{a}{s(s+a)}$	$\frac{1}{z-1}-\frac{e^{-amT}}{z-e^{-aT}}$
$\frac{a}{s^2(s+a)}$	$\frac{mT-1/a}{z-1}+\frac{T}{(z-1)^2}+\frac{e^{-amT}}{a(z-e^{-aT})}$
$\frac{a^2}{s(s+a)^2}$	$\frac{1}{z-1}-e^{-amT}\frac{1+amT}{z-e^{-aT}}+\frac{aTe^{-aT}}{(z-e^{-aT})^2}$
$\frac{a}{s^2+a^2}$	$\frac{z\sin amT+\sin(1-m)aT}{z^2-2z\cos aT+1}$, $\Delta(1-m)$, $a\pm\frac{k\pi}{T}$, $k=1, 2, 3\ldots\ldots$
Special case in which $a=k\pi/T$, $k=1, 2, 3, \ldots$	
$\frac{\frac{k\pi}{T}}{s^2+\left(\frac{k\pi}{T}\right)^2}$	$\frac{\sin mk\pi}{z+(-1)^{k+1}}$
$\frac{\pi/T}{s^2+(\pi/T)^2}$	$\frac{\sin m\pi}{z+1}$
$\frac{s}{s^2+a^2}$	$\frac{\cos amT-\cos(1-m)aT}{z^2-2z\cos aT+1}$
$\frac{z+a}{(s+a)^2+\beta^2}$	$\frac{[(a-a)^2+\beta^2]^{1/2}}{\beta}e^{-amT}\left[\frac{e^{-aT}\sin(\beta T-m T\beta-\theta)+z\sin(mT+\theta)}{z^2-2ze^{-aT}\cos\beta T+e^{-2aT}}\right]$, $\theta=\tan^{-1}\left(\frac{\beta}{a-a}\right)$

* For more extended tables, see reference 1.

If there exists a real, pure delay in the system which is an integer number of the sampling instants, this can be included in the transfer function $G(s)$.

EXAMPLE 1

Suppose

$$G(s)=\frac{1}{s+a} \tag{5}$$

then

$$G^*(s)=\frac{1}{1-e^{-aT}z^{-1}}=\frac{z}{z-e^{-aT}} \tag{6}$$

and

$$G^*(z, m)=z^{-1}\frac{e^{-maT}}{1-e^{-aT}z^{-1}}=\frac{e^{-maT}}{z-e^{-aT}}(\Delta>0) \tag{7}$$

In Table I a few pairs of the modified pulsed transfer function are inserted which cover most of the encountered cases.

To find the modified z-transform of the response of a system shown in Fig. 1, the usual z-transform multiplication is applied

$$C^*(z, m)=R^*(z)\times G^*(z, m) \tag{8}$$

and

$$C(n, m)T=\frac{1}{2\pi i}\int_\Gamma R^*(z)\times G^*(z, m)\times z^{n-1}dz \tag{9}$$

where Γ encloses all singularities of the integrand in equation 9.

In effecting the integration m is considered constant and the use of Table I of reference 5 can be applied. However, to obtain the time response, let m vary from zero to unity, which yields all the information between sampling instants.

Suppose a step input is applied to the system just described. Then

$$C^*(z, m)=R^*(z)\times G^*(z, m) \tag{10}$$

$$C(n, m)T=\frac{1}{2\pi i}\int_\Gamma\frac{z}{z-1}\frac{e^{-maT}}{z-e^{-aT}}z^{n-1}dz=e^{-maT}\frac{1-e^{-aTn}}{1-e^{-aT}}(\Delta>0) \tag{11}$$

The plot of this function for $a=1$, $T=1$ is shown in Fig. 2. This response may also be obtained with the division method,[3] which agrees with the results obtained.

Application to Feedback Systems

The modified pulsed transfer function can be applied for feedback systems. Imagine the system of Fig. 3 to be replaced by that of Fig. 4 in which there is a fictitious delay in the forward path and an equal but negative delay in the feedback path. The performance will be identical with that of the original system, except that the calculated output sequence will refer to the time $t=(n-1+m)T$.

With reference to Fig. 3, the over-all pulsed transfer function is

$$\frac{C^*(z)}{R^*(z)}=\frac{G^*(z)}{1+G^*(z)} \tag{12}$$

and in Fig. 4

$$\frac{C^*(z, m)}{R^*(z)}=\frac{G^*(z, m)}{1+G^*(z)} \tag{13}$$

If there exists a real delay within the servo loop, it must be included in the over-all transfer function. Furthermore, if there exist components in the feedback path, then equation 13 will become

$$\frac{C^*(z, m)}{R^*(z)}=\frac{G^*(z, m)}{1+HG^*(z)} \tag{14}$$

Thus it is noticed that the modified z-transform method can be conveniently applied for synthesis procedures, for it can be easily consolidated with the z-transform method.

Because one can use the modified z-transform, letting m vary between zero and unity, this is a convenient method for synthesis procedures of sampled-data control systems. Before this method is used, one point must be indicated: when $m=1$, $G^*(z, m)$ approaches $G^*(z)$ provided that the impulse response $\mathcal{L}^{-1}[G(s)]$, of the system is continuous over the range of variation of m and through unity, as in most cases of the servo systems. Where this condition is not fulfilled, the indication $\Delta>0$ should be indicated. A case of this type is thoroughly discussed in the preceding example.

In comparing equation (13) with equation (12), the following observations can generally be deduced:

1. The poles of $C^*(z, m)$ are the same as the poles of $C^*(z)$, provided there is no cancellation of a pole and zero of $G^*(z)$.
2. The zeros of $C^*(z, m)$, in number and magnitude (or location in the z-plane), are different from the zeros of $C^*(z)$.

The fact that only the zeros of $C^*(z, m)$ are different from the corresponding zeros of $C^*(z)$ considerably facilitates the design procedures using the modified z-transform method.

To obtain the actual response $C(n, m)T$, the following contour integral may be used

$$C(n, m)T=\frac{1}{2\pi i}\int_\Gamma\frac{G^*(z, m)}{1+G^*(z)}\times R^*(z)z^{n-1}dz \tag{15}$$

where Γ is a path of integration in the z-plane that includes all the poles of the

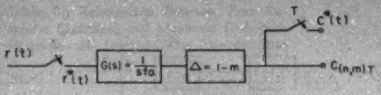
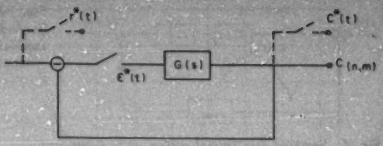


Fig. 1. Sampled-data system

Fig. 3 (right). Sampled - data control system



integral in equation 15. This integral can be easily evaluated with the use of tables[5] for m is constant in the integration; and to cover the whole response, m continually varies from zero to unity and when m equals unity one obtains $C(nT)$, the response at the sampling instants.

In actual design of sampled-data control, one may not need to evaluate the response for all instants of time. The overshoot and the peak time are the most important quantities for design procedures. These quantities can be obtained using the following procedure. From equation 15 one obtains $C(n, m)T$, and when m is unity, the response is $C(nT)$.

To find the maximum of $C(nT)$, one may apply the method indicated in references 5 and 6, which will yield $C(nT)_{max}$, and n_{max}, which in this case is integer. Once n_{max} is obtained, it can be inserted in equation 15 which will yield $C(n_{max}, m)T$.

The function $C(n_{max}, m)$ is continuously varying with m, and the laws of ordinary differentiation can be applied, i.e., to find m for the maximum, one may use the following equation

$$\frac{\partial C(n_{max}, n)}{\partial m} = 0 \tag{16}$$

The solution of equation 16 yields the value of m for the maximum in the actual response. Alternately, the maximum time can be obtained, and so can the maximum overshoot.

The procedure is a straightforward one and involves the ordinary differentiation, as encountered in the continuous case. However, this is based on the assumption that the maximum overshoot occurs near the maximum of the sampling instant, as is usually the case in most orthodox sampled-data control systems.

In case the maximum is not near the maximum of the response at sampling instants, as in the case of hidden oscillations, one may apply the following alternate procedure. (The maximum in these rare cases generally happens in the first sampling interval because the impulsive response is maximum at these intervals if the system is stable; see the Appendix.) Differentiate equation 15 to m and equate to zero.

$$\frac{\partial C(n, m)T}{\partial m} = 0 \tag{17}$$

The solution of equation 17 yield a value of m_{max} as a function of (n). This value of m_{max} is inserted in equation 15 to yield $C(n, m_{max})T$.

Since it is known that m varies only between zero to unity, equation 17 can be solved for n, which yields the values of m indicated. So, both n_{max} and m_{max} can be found and the maximum of the total response.

As a final case, it might happen that the actual maximum is equal to $C(n_{max})$, then this case can be detected by finding that $C(n_{max}, m) = C(n_{max})$ for all values of m.

EXAMPLE 2

To illustrate the procedure outlined previously, its application to the simple feedback system shown in Fig. 4 is described. The forward transfer function is

$$G(z) = \frac{1}{s(s+1)} \tag{18}$$

From tables

$$G^*(z) = \frac{(1-\epsilon^{-T})z}{(z-1)(z-\epsilon^{-T})} \tag{19}$$

From Table I

$$G^*(z, m) = \frac{z-\epsilon^{-T}-\epsilon^{-Tm}(z-1)}{(z-1)(z-\epsilon^{-T})} \tag{20}$$

The output $C^*(z, m)$ is given for a step input as

$$C^*(z, m) = R^*(z) \times \frac{G^*z, m)}{1+G^*(z)} \tag{21}$$

while

$$R^*(z) = \frac{z}{z-1} \tag{22}$$

then

$$C^*(z, m) \text{ for } T = 1n\ 2 \text{ or } \epsilon^{-T} = \frac{1}{2} \tag{23}$$

$$C^*(z, m) = \frac{z}{z-1} \frac{1-\left(\frac{1}{2}\right)^m z+\left(\frac{1}{2}\right)^m-\frac{1}{2}}{(z-1)\left(z-\frac{1}{2}\right)+\frac{1}{2}z} \tag{24}$$

Letting

$$\left(1-\frac{1}{2}\right)^m = b, \left(\frac{1}{2}\right)^m-\frac{1}{2} = a \tag{25}$$

gives

$$C^*(z, m) = \frac{z}{z-1} \frac{bz+a}{\left(z-\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^2} \tag{26}$$

Applying the inversion equation or tables, to obtain

$$C(n, m)T = 1+\frac{2}{\sqrt{2}n}\left[\left(\frac{1}{2}\right)^{2m}-\left(\frac{1}{2}\right)^m+\frac{1}{2}\right]^{\frac{1}{2}} \sin\left(n\frac{\pi}{4}+\frac{\pi}{4}+\pi+\tan^{-1}(2^m-1)\right) \tag{27}$$

for $m = 1$, the response will be at the sampling instants only

$$C(nT) = 1-\frac{1}{\sqrt{2}n}\sin\left(n\frac{\pi}{4}+\frac{\pi}{4}+\tan^{-1}1\right) \tag{28}$$

For the maximum n, the following condition should be satisfied[5]

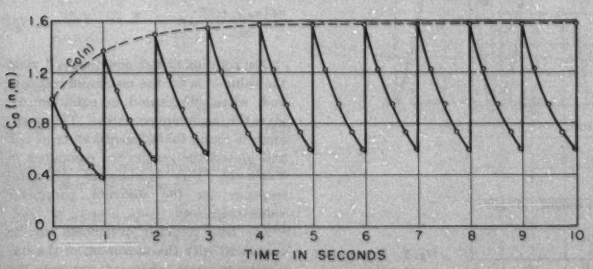


Fig. 2. Output response of a system in example 1

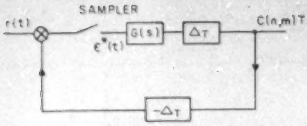


Fig. 4. Sampled-data control system with a fictitious delay

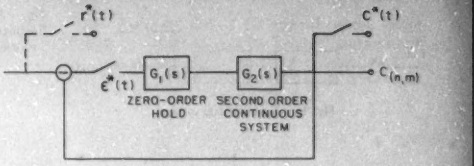


Fig. 6 (right). Second - order sampled - data control system

$$n\frac{\pi}{4}+\frac{\pi}{4}=\pi \tag{29}$$

or $n=3$ (30)

and $n_{max}=3+1=4$ (31)

Substituting for n_{max} in equation 28, to obtain

$$C(n_{max})T=1-\frac{1}{4}\sin\frac{6\pi}{4}=1.25 \tag{32}$$

Now, to find the maximum response, obtain $C(n_{max}, m)T$ as follows:

$$C(n_{max}, m)T=1-\frac{2}{4}\left[\left(\frac{1}{2}\right)^{2m}-\left(\frac{1}{2}\right)^{m}+\frac{1}{2}\right]^{\frac{1}{2}} \sin\left[\pi+\frac{\pi}{4}+\tan^{-1}(2^{m}-1)\right] \tag{33}$$

This equation can be reduced to

$$C(n_{max}, m)T=1+\frac{1}{4}=1.25 \text{ for all } m \tag{34}$$

Thus it is noticed that the actual maximum is the same as the maximum response at the sampling instants. The actual plot of equation 27 for all values of m is seen in Fig. 5.

Application to Second-Order Systems

For a second-order system such as is shown in Fig. 6, $G_2(s)$ equals

$$G_2(s)=\frac{\omega_0^2}{s(s+2\omega_0\eta)} \tag{35}$$

where

ω_0 = undamped natural frequency of the second-order system
η = damping ratio of the system

With the use of the zero-order hold function whose transfer function is

$$G_1(s)=\frac{1-e^{-Ts}}{s} \tag{36}$$

then the z-transform of $G_1(s)G_2(s)$ may be found[3] with the use of Table 1 to yield

$$G^*(z)=\frac{\omega_0}{2\eta}\frac{T}{z-1}-\frac{1}{4\eta^2}\frac{1-e^{-2\omega_0\eta T}}{z-e^{-2\omega_0\eta T}} \tag{37}$$

Similarly from Table I, the modified z-transform of the transfer function may be obtained as follows

$$G^*(z, m)=\frac{\omega_0}{2\eta}\frac{z-1}{z}\times\left[\frac{T}{(z-1)^2}+\frac{mT}{z-1}-\frac{1}{2\omega_0\eta}\left(\frac{1}{z-1}-\frac{e^{-2\omega_0\eta Tm}}{z-e^{-2\omega_0\eta T}}\right)\right] \tag{38}$$

It is noticed from equation 38 that when m is unity, $G^*(z, m)=G^*(z)$, the z-transform of the transfer function.

The output-input function can be written as

$$\frac{C^*(z, m)}{R^*(z)}=\frac{G^*(z, m)}{1+G^*(z)} \tag{39}$$

For step input

$$R^*(z)=\frac{z}{z-1} \tag{40}$$

Thus

$$C^*(z, m)=\frac{z}{z-1}\frac{G^*(z, m)}{1+G^*(z)} \tag{41}$$

To obtain the time response, one may use the inversion equation, which yields

$$C(n, mT)=\frac{1}{2\pi i}\int_\Gamma \frac{z}{z-1}\frac{G^*(z, m)}{1+G^*(z)}z^{n-1}\,dz \tag{42}$$

This integral can be evaluated from tables in reference 5, for m is constant in the integration, and when m is unity one obtains $C(nT)$.

To obtain the maximum of the response in this case, solve for n_{max} using the method indicated in references 5 and 6. Then substitute n_{max} in equation 42 to obtain $C(n_{max}, m)T$. To obtain the maximum of this response, the following equation is equated to zero.

$$\frac{\partial C(n_{max}, m)T}{\partial m}=0 \tag{43}$$

The solution of this equation will yield the value of m for the maximum response and, when substituted in equation 42, gives the maximum response. This maximum is found for different values of $\omega_0 T$ and η, and the results are computed in Table II. Fig. 7 indicates the total response for the different parameter values calculated.

The time response overshoots can be correlated with the maximum of the frequency response when z transverses the unit circle, i.e.[5]

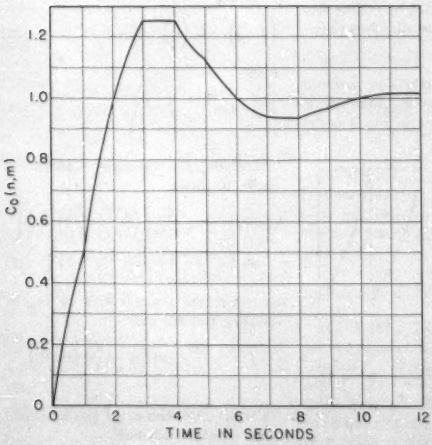


Fig. 5. Response of a system in example 2

Table II. Relationship Between Time Response Overshoot and Maximum Frequency Response of Second-Order Sampled-Data Control System

$\omega_0 T$	η	τ_{max} in Sampling Instant τ_p	M_{max}	M_p, Per Cent Overshoot
1	0.3	3.35	4.85	82
1	0.5	3.45	1.96	45
1	0.8	4.05	1.12	15
1	1.0	4.85	1.01	5

$$M_{max} = \max \frac{G^*(z)}{1+G^*(z)}\bigg|_{z \to e^{j\omega}} \qquad (43A)$$

This correlation between M_{max} and maximum overshoot are tabulated for $\omega_0 T = 1$ and for various values of η as in Table II.

It is evident that it is not necessary to find the maximum of $[G^*(z, m)]/[1+G^*(z)]$ because for each form of $G(z)$ corresponds to only one form of $G^*(z)$ and/or $G^*(z, m)$. Thus the maximum of $G^*(z)/[1+G^*(z)]$ is sufficient.

In conclusion, one might mention that the limitations of the z-transform method imposed previously is removed, for one can obtain the correlation between the frequency response and the maximum of the transient not only at the sampling instance, as has been done previously,[8] but for the total response as well. This correlation can be used advantageously for the design of the sampled-data control system.

Although the correlation obtained is only for second-order systems, it can be modified for higher order systems when dominance holds, as shown in reference 7.

EXAMPLE 3

In certain cases the maximum response coincides with the maximum of the sampling response, in which case $C(n_{max}, m) = C(n_{max})$. As an example, let

$$G(z) = \frac{13.4}{s(s+10)} \qquad (44)$$

The corresponding $G^*(z)$ is, for $T = 0.1$ second

$$G^*(z) = \frac{0.846z}{z^2 - 1.368z + 0.368} \qquad (45)$$

and the modified z-transform $G^*(z, m)$, from Table I, is as follows:

$$G^*(z, m) = 1.34\ \frac{z - e^{-1} - e^{-m}(z-1)}{(z-1)(z-e^{-1})} \qquad (46)$$

The output transform for step input equals

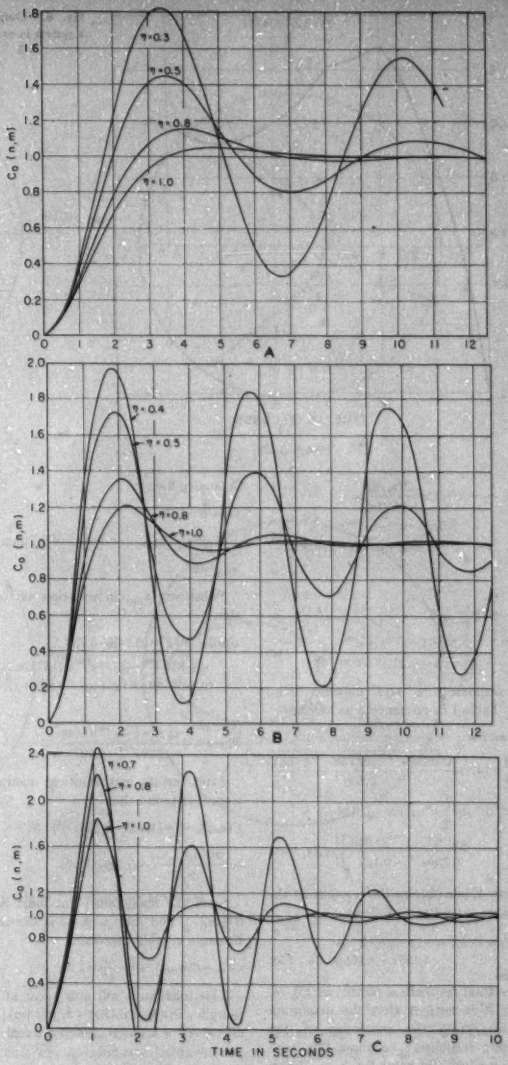


Fig. 7. System responses for a step input

A—$\omega_0 T = 1$
B—$\omega_0 T = 2$
C—$\omega_0 T = 4$

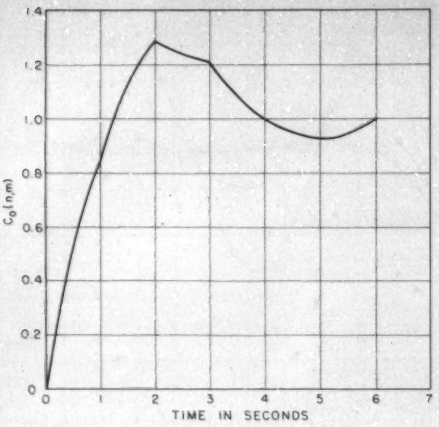


Fig. 8. Output of a system in example 3

$$C^*(z, m) = R^*(z) \times \frac{G^*(z, m)}{1+G^*(z)} = \frac{z}{z-1} \frac{1.34[z-\epsilon^{-1}-\epsilon^{-m}(z-1)]}{z^2-1.368z+0.368+0.846\epsilon} \quad (47)$$

or

$$C^*(z, m) = \frac{z}{z-1} \times \frac{1.34[z(1-\epsilon^{-m})+\epsilon^{-m}-\epsilon^{-1}]}{[(z-0.261)^2+0.3]} \quad (48)$$

The response $C(n, m)T$ may be found from Table I in reference 5 as follows

$$C(n, m)T = \left\{0.747+\frac{\sqrt{0.846\epsilon^{-2m}-0.758\epsilon^{-m}+0.311}}{0.503}(0.606)^n \sin\left[n \times 1.126+0.638+\pi+\frac{\tan^{-1}(1-\epsilon^{-m})0.548}{0.739\epsilon^{-m}-0.107}\right]\right\}1.34 \quad (49)$$

For $m=1$, one obtains the response at the sampling instants, as follows

$$C(nT) = [0.747-0.763(0.606)^n \sin (1.126n+1.765)]1.34 \quad (50)$$

The total response is plotted in Fig. 8, where it is noticed that the maximum coincides with the maximum of the sampling response. To show this condition analytically, solve for n_{max}. From equation 50, the maximum is found by solving the following equation:[6]

$$n\theta+\lambda = k\pi \quad (51)$$

Since

$$\theta = 1.1264 \quad (52)$$
$$\lambda = 1.126^\circ$$

Assuming $k=1$ (53)

thus $n=1.79$ (54)

Taking the upper value of n, then

$$n_{max}=2 \quad (55)$$

Substitute n_{max} in equation 49 to obtain

$$C(n_{max}, m)T = [0.747-0.732\sqrt{0.846^{-2m}-2.758\epsilon^{-m}+0.711} \times (0.368) \sin (1.1264 \times 2+0.638+\lambda)]1.34 \quad (56)$$

where $\lambda = \frac{\tan^{-1}(1-\epsilon^{-m})0.548}{0.739\epsilon^{-m}-0.107}$

After certain mathematical manipulation, it is found that

$$C(n_{max}, m) = (1.155-0.522\epsilon^{-m})1.34 \quad (57)$$

or $$\frac{\partial C(n_{max}, m)}{\partial m} = 0.699\epsilon^{-m} \neq 0 \quad (58)$$

Since the maximum point must lie in $0 \leqq m \leqq 1$, and since $C(n, m)$ is increasing throughout this interval

$$C_{max} = C(n_{max}), \text{ when } m=1 \quad (59)$$

This maximum will also occur at the sampling instants if $G(s) = k/[s(s+a)]$, for all values of k and a, which are real and positive.

Root-Locus Approach

It has been indicated in reference 7 that sampled-data systems can be analyzed and designed using the root-locus approach. It is shown that when the system is stable, the roots of $1+G^*(z)$ should lie inside the unit circle. In this work, constant overshoot plots have been computed which would determine the parameter values to be used for certain overshoot and peak times.

For the second-order system whose $G(s)$ is equal to

$$G(s) = \frac{\omega_0^2}{s^2(s+2\omega\eta)}(1-\epsilon^{-Ts}) \quad (60)$$

one can form the input-output z-transform relationship, from the table in reference 5, thus

$$\frac{C^*(z)}{R^*(z)} = \frac{G^*(z)}{1+G^*(z)} \quad (61)$$

For step input

$$R^*(z) = \frac{z}{z-1} \quad (62)$$

the output z-transform is equivalent to

$$C^*(z) = \frac{z}{z-1} \frac{G^*(z)}{1+G^*(z)} \quad (63)$$

The locus of the roots of $1+G^*(z)$ is plotted in Fig. 9 and computed in Table III. For this locus, $\omega_0 T$ is held constant and η is variable, which will yield the corresponding overshoot. It is noticed from the plot that when the roots are near the origin, the system will be more stable.

Constant Overshoot Loci in the z-Plane

The advent of the root-locus method facilitates the analysis and synthesis of sampled-data control systems. To obtain information on the system response, constant overshoot loci can be plotted inside the unit circle in the z-plane which serves as design charts for sampled systems.

These constant overshoot loci are calculated for second-order sampled-data control systems, i.e.,

$$G(s) = \frac{\omega_0^2(1-\epsilon^{-Ts})}{s^2(s+2\omega\eta)} \quad (64)$$

The input-output z-transform is found with the use of tables, as follows

$$\frac{C^*(z, m)}{R^*(z)} = \frac{G^*(z, m)}{1+G^*(z)} \quad (65)$$

For step input $R^*(z) = \frac{z}{z-1}$

Thus $$C^*(z, m) = \frac{z}{z-1} \frac{G^*(z, m)}{1+G^*(z)} \quad (66)$$

The response equals

$$C(n, m)T = \frac{1}{2\pi i} \int_\Gamma \frac{z}{z-1} \frac{G^*(z, m)}{1+G^*(z)} z^{n-1}\, dz \quad (67)$$

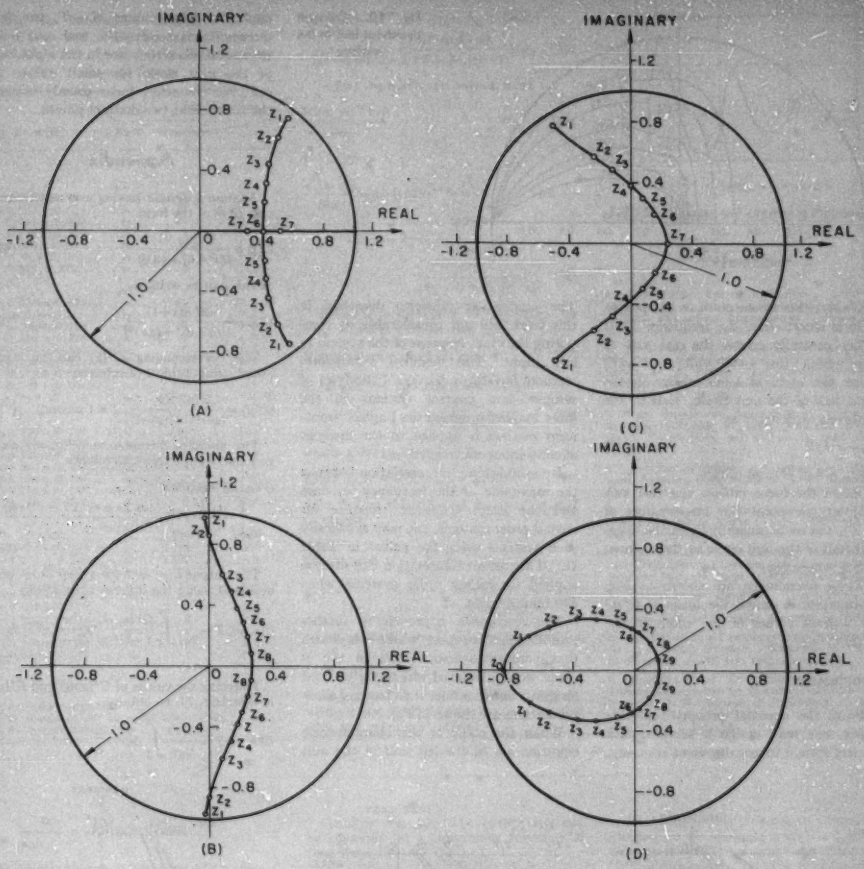


Fig. 9. Loci of roots in z-plane

A—$\omega_0 T = 1$
B—$\omega_0 T = 2$
C—$\omega_0 T = 3$
D—$\omega_0 T = 4$

The maximum of the response can be found using the method indicated previously. It is shown that to keep the overshoot constant, both ($\omega_0 T$) and η must both be varied.

These constant overshoot loci are plotted in Fig. 10 where it is noticed that the roots of $1+G^*(z)$ can be real or imaginary depending on the values of $\omega_0 T$ and η.

It can be shown that these constant overshoot loci can be used for higher systems when dominance holds. Thus, these loci are important in the design of sampled-data control systems.

Fig. 10 indicates the constant overshoot loci for the total response; these constant loci are for various values of overshoot and peak time. It is noticed also from Fig. 10 that when the system has no overshoot, the roots should be on the positive real axis in the z-plane.

Correlation Between Root-Locus and Frequency Locus

The frequency locus $G^*(z)$ is the locus of all the points of $G^*(z)$ when z traverses the unit circle in the z-plane. The correlation between the root-locus and the frequency locus is discussed in detail in reference 7. However, in this paper it is shown that for the second-order system, the frequency locus will cross the negative

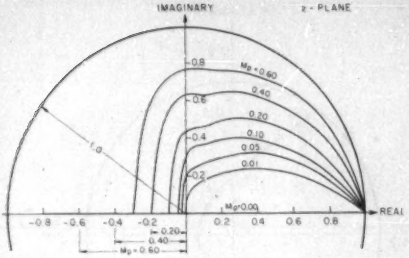


Fig. 10. Constant overshoot loci in the z-plane

real axis either at one point or two points.

It is shown that the frequency locus $G^*(z)$ generally crosses the real axis at two points (for small values of $\omega_0 T$) when the roots of $1+G^*(z)$ are in the right half of the unit circle; that is, the roots can be denoted as

$$z_{1,2} = A \pm jB \tag{68}$$

Both A and B are positive.

When the locus crosses the real axis only at one point (for large values of $\omega_0 T$), the roots generally lie in the negative half of the unit circle in the z-plane; i.e., A is negative.

These phenomena are shown in Fig. 11 in which is shown the locus of $G^*(z)$ for various values of the roots of $1+G^*(z)$.

Conclusions

From the material presented in this paper, one may design a sampled-data control system to any degree of accuracy. The modified z-transform discussed in this work will aid considerably in computing the exact response of the system to any input. The modified z-transform method facilitates the exact analysis of sampled-data control systems in the same manner in which the Laplace transform method is applied to the analysis of continuous control systems.

By establishing the correlation between the maximum of the frequency-response and the exact transient response for second-order systems, one may synthesize such systems using the values in Table II. This design information can also be applied for higher order systems when dominance holds.

The root-locus approach is further extended in this report, where it is shown that constant-overshoot loci in the z-plane were calculated which can be used for design information, and charts of those tabulations are shown in Fig. 10.

When the roots of the characteristic equation are in the left half of the unit circle for large values of $\omega_0 T$, the frequency locus crosses the real axis only once, and when they are in the right half of the unit circle for small values of $\omega_0 T$, the frequency locus usually crosses the real axis at two distinct points.

Appendix

Assume a system having a transfer function $G(s)$ of the form

$$G(s) = \frac{27.8(s+1)}{s(s^3+6s+48.6)} \tag{69}$$

or this can be written

$$G(s) = \frac{27.8(s+1)}{s[(s+3)^2+(2\pi)^2]} \tag{70}$$

The corresponding $G^*(z)$ can be computed using Table I of reference 5 as

$$G^*(z) = \frac{0.543 \times z}{(z-0.06)(z-1)}, \quad T = 1 \text{ second} \tag{71}$$

The modified z-transform $G^*(z, m)$ may be found from Table I, as follows

$$G^*(z, m) = 0.572 \times \left[\frac{1}{z-1} - \frac{e^{-2m}(\cos 2\pi m - 7.27 \sin 2\pi m)}{z-e^{-1}}\right] \tag{72}$$

The output $C(n, m)T$ for a step input can be found using the inverse z-transform

$$C(n, m)T = \frac{1}{2\pi i}\int_\Gamma \frac{G^*(z, m)}{1+G^*(z)} \times \frac{z}{z-1} z^{n-1}\, dz \tag{73}$$

Inserting the values of $G^*(z,m)$ and $G^*(z)$ in equation 74, to obtain

$$C(n, m)T = \frac{0.572}{2\pi i}\int_\Gamma \times$$

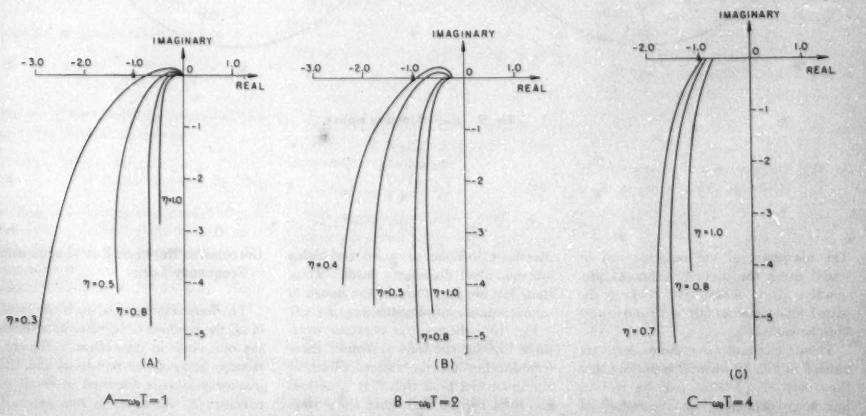


Fig. 11. Frequency loci in G*(z)-plane

Table III. Relationship Between Location of Roots and Time Sequence Response Overshoot for Second-order Sampled-Data Control System

Location	Root	M_p, Per Cent Overshoot	n
$\omega T=1.0$			
$0.568\pm j0.752$	z_1	82	0.3
$+0.500\pm j0.618$	z_2	40	0.5
$0.444\pm j0.436$	z_3	15	0.8
$0.426\pm j0.320$	z_4	5	1.0
$0.416\pm j0.195$	z_5	0.37	1.2
0.412	z_6	0	1.35
0.513; 0.310	z_7 z_8	0	1.4
$\omega T=2$			
$-0.026\pm j0.971$	z_1	94	0.4
$\pm 0.854j$	z_2	73	0.5
$0.083\pm j0.598$	z_3	35	0.8
$0.132\pm j0.478$	z_4	20	1.0
$0.174\pm j0.380$	z_5	8	1.2
$0.208\pm j0.291$	z_6	4	1.4
$0.237\pm j0.205$	z_7	0.62	1.6
$0.261\pm j0.091$	z_8	0*	1.8
$+0.27$	z_9	0	1.9005
$\omega T=3$			
$-0.500\pm j0.795$	z_1	105	0.5
$-0.240\pm j0.568$	z_2	49	0.8
$-0.124\pm j0.482$	z_3	23	1.0
$-0.003\pm j0.394$	z_4	15	1.3
$+0.080\pm j0.302$	z_5	7	1.6
$0.156\pm j0.195$	z_6	0.85	2.0
0.299; 0.190	z_7	0.00	2.4
$\omega T=4$			
$-0.818\dagger$	z_1	222	0.6
$-0.673\pm j0.222$	z_2	148	0.7
$-0.554\pm j0.283$	z_3	122	0.8
$-0.375\pm j0.330$	z_4	84	1.0
$-0.246\pm j0.336$	z_5	49	1.2
$-0.111\pm j0.314$	z_6	22	1.5
$-0.017\pm j0.277$	z_7	8	1.8
$+0.031\pm j0.248$	z_8	0	2.0
$+0.105\pm j0.180$	z_9	1	2.4
$+0.159\pm j0.082$	z_{10}	0‡	2.8
0.190; 0.154	z_{11}	0	2.92

* At integer instant only.
† Convergence not apparent until $n=26$.
‡ At integer instant only.

$$z-0.05-e^{-2m}(z-1)(\cos 2\pi m-\frac{7.27\sin 2\pi m)}{[(z-0.05)(z-1+0.543z)](z-1)}z^n dz \quad (74)$$

To find the maximum output, form

$$\frac{\partial C(n,m)T}{\partial m}=0 \quad (75)$$

which gives

$$\frac{\partial C(n,m)}{\partial m}=0=0.572e^{-1.2m}(-15.53\sin 2\pi m+48.6\cos 2\pi m)\times\frac{1}{2\pi i}\int_\Gamma\times\frac{z^n dz}{\Gamma(z^2-0.507z+0.05)} \quad (76)$$

Since the integral is never zero except for $n=0$, which is a trivial case

$$48.6\cos 2\pi m-15.53\sin 2\pi m=0 \quad (77)$$

or

$$\sqrt{48.6^2+15.53^2}\cos(2\pi m+\theta)=0,$$

$$\tan\theta=\frac{15.53}{48.6},\ \theta=17.8^\circ \quad (78)$$

which gives

$$2\pi m+17.8^\circ=\pi/2,\ m_{max}=0.2005 \quad (79)$$

Substituting this value of m_{max} in $C(n,m)T$ to obtain

$$C_{max}(n,m)T=\frac{1}{2\pi i}\int_\Gamma\times\frac{(4.63z-3.68)z^n dz}{(z-1)(z-0.135)(z-0.372)} \quad (80)$$

which yields for $n=1$, $C_{max}=2.64$; and for $n=2$, $C_{max}=1.885$;
for $n=3$; $C_{max}=1.37$ (81)

The plot of $C(n,m)T$ is shown in Fig 12, where it is noticed that the maxima coincide with the calculated values.

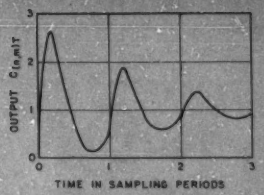


Fig. 12. Response of a system in Appendix

References

1. The Pulse Transfer Function and Its Application to Sampling Servosystems, R. H. Barker. *Proceedings*, Institution of Electrical Engineers, London, England, vol. 99, pt. IV, 1952.

2. Extension of Conventional Techniques to the Design of Sampled-Data Systems, W. K. Linvill, R. W. Sittler. *IRE Convention Record* Institute of Radio Engineers, New York, N. Y., pt. I, 1953, pp. 99–104.

3. The Design of Sampled-Data Feedback Systems, Gladwyn V. Lago, John G. Truxal. *AIEE Transactions*, vol. 73, pt. II, Nov. 1955, pp. 247–53.

4. Analysis of Errors in Sampled-Data Feedback Systems, Jack Sklansky, J. R. Ragazzini. *Ibid.*, vol. 74, pt. II, May 1955, pp. 65–71.

5. Analysis and Synthesis of Sampled-Data Control Systems, Eliahu I. Jury. *Ibid.*, vol. 73, pt. I, Sept. 1954, pp. 332–46.

6. The Effect of Pole and Zero Locations on the Transient Response of Sampled-Data Systems. *Ibid.*, vol. 74, pt. II, Mar. 1955, pp. 41–48.

7. Correlation Between Root-Locus and Transient Response of Sampled-Data Control Systems. *Ibid.*, vol. 75, pt. II, Jan. 1956, pp. 427–36.

8. Additions to z-Transformation Theory for Sampled-Data Systems, Gladwyn V. Lago. *Ibid.*, vol. 74, pt. II, Jan. 1955, pp. 403–08.

9. Extension of Continuous-Data System Design Techniques to Sampled-Data Control Systems, G. W. Johnson, D. P. Lindorff, C. G. A. Nordling. *Ibid.*, Sept. 1955, pp. 252–63; discussion, Eliahu I. Jury, pp. 259–60.

Discussion

G. Farmanfarma (University of California, Berkeley, Calif.): The modified z-transform discussed provides an exact method for the analysis and synthesis of sampled-data control systems between sampling instants. As the author points out, it can also furnish the value of the output at the sampling instants if the impulse response $g(t)=\mathcal{L}^{-1}G(s)$ is continuous for all $0\leq t\leq\infty$. This should provide the design engineer with the necessary tool to study his designs more critically, thus eliminating systems that may have high overshoots between the sampling periods, even though they have a reasonable response at the sampling instants. It would be unfortunate if the application of the modified z-transform were considered only as related to sampled-data control systems, for this would mean neglect of its abstract mathematical nature. It is the purpose of this discussion to point out a useful application of the modified z-transform, that of evaluating the inverse Laplace of certain types of functions.

Consider Fig. 13. The output $c(t)$ can be obtained as a continuous function of time from $C(s)$

$$c(t)=\mathcal{L}^{-1}C(s)=\mathcal{L}^{-1}R^*(s)G(s) \quad (82)$$

However, by applying the modified z-transform, $c(t)$ can be expressed at any instant of time $t=(n-1+m)T$, as the equation

$$c(t)=z_m{}^{-1}C(z,m)=z_m{}^{-1}[R(z)G(z,m)]_{t=(n-1+m)T} \quad (83)$$

$0<m<1$ or $0\leq m\leq 1$

Since the solution is unique, by equating equations 82 and 83, a general relation between the inverse Laplace and the inverse modified z-transform of a function is obtained.

Thus

$$c(t)=\mathcal{L}^{-1}C(s)=z_m{}^{-1}C(z,m)_{t=(n-1+m)T} \quad (84)$$

where $C(z,m)$, the modified z-transform of $C(s)$, is given by equation 4.

The relation (84) provides an alternative method for evaluating the inverse Laplace through the use of the modified z-transform. In the case when $C(s)$ is a polynomial in s, it is simpler to use the ordinary methods, and evaluate $c(t)$ directly from $C(s)$. But in some cases when $C(s)$ is the Laplace transform of a periodic function, i.e., of the form $(1/1-e^{-Ts})R(s)$ or has factorable terms in the form of e^{aTs}, then it becomes advantageous to obtain the inverse time function through the use of the modified z-transform. To illustrate, consider the example

$$C(s)=\frac{1}{1-e^{-Ts}}\frac{1}{s+a} \quad (85)$$

the inverse time function is expressed as

$$c(t)=\frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty}\frac{e^{st}}{(1-e^{-Ts})(s+a)}ds \quad (86)$$

the integrand has simple poles at $s=-a$ and $s=\pm j2\pi k/T$, $k=0,1,2,\ldots$ evaluating the integral at these simple poles gives the following:

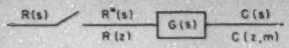


Fig. 13. Sampled-data system

$$c(t)=\frac{e^{-at}}{1-e^{aT}}+\frac{1}{aT}+\frac{2}{T}\sum_{k=1}^{\infty}\times \frac{a\cos\frac{2\pi k}{T}t+\frac{2\pi k}{T}\sin\frac{2\pi k}{T}t}{a^2+\left(\frac{2\pi k}{T}\right)^2} \quad (87)$$

With the use of equation 84 the inverse time function may also be written as

$$c(t)=\mathcal{L}^{-1}\,C(s)=z_m{}^{-1}[C(z,m)]_{t=(n-1+m)T} \quad (88)$$

where

$$C(z,m)=z_m\left[\frac{1}{1-e^{-Ts}}\frac{1}{(s+a)}\right]$$

$$C(z,m)=\frac{z}{z-1}\frac{e^{-amT}}{z-e^{-aT}} \quad 0<m<1 \quad (89)$$

From equation 9

$$c(n-1+m)T=\frac{1}{2\pi j}\int_{\tau}\frac{z^n}{z-1}\frac{e^{-amT}}{z-e^{-aT}}dz \quad (90)$$

This yields

$$c(n-1+m)T=\frac{e^{-amT}-e^{-aT(n+m)}}{1-e^{-aT}} \quad (91)$$

multiplying equation 91 by e^{aT} and substituting $t=(n-1+m)T$ gives

$$c(t)=\frac{e^{-amT}}{1-e^{-aT}}-\frac{e^{aT(1-m)}}{1-e^{aT}} \quad (92)$$

The advantage of using the modified z-transform to obtain the inverse Laplace transform in this case is evident from the simplicity of equation 92 as compared to equation 87. It is seen that the two solutions have the same transient term. This is expected if they are to be equivalent at all values of time.

Equating the steady-state component gives

$$\frac{e^{aT(1-m)}}{e^{aT}-1}=\frac{1}{aT}+\frac{2}{T}\sum_{k=1}^{\infty}\times$$

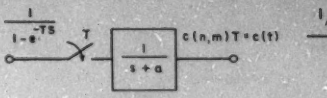


Fig. 15. Sampled-data system

$$\frac{a\cos\frac{2\pi k}{T}t+\frac{2\pi k}{T}\sin\frac{2\pi k}{T}t}{a^2+\left(\frac{2\pi k}{T}\right)^2} \quad (93)$$

from which, by substitution of $t=(n-1+m)T$, is obtained the relation

$$\sum_{k=1}^{\infty}\frac{a\cos 2\pi km+\frac{2\pi k}{T}\sin 2\pi km}{a^2+\left(\frac{2\pi k}{T}\right)^2}= \frac{T[e^{aT(1-m)}]}{2(e^{aT}-1)}-\frac{1}{2a} \quad (94)$$

$0<m<1$

Equation 94 is a by-product of the relation 84. The right hand side is the closed form of the infinite series on the left-hand side. This identity has been checked and found to be correct for various values of a, T, and m.

In general, if $C(s)$ is the Laplace transform of a periodic function, the inverse Laplace is always in the form of an infinite series, but the use of relation 84 enables one to find a closed form for this infinite series; thus there is no reason why $C(s)$ should not be synthesized to give an infinite series whose sum is to be found in the time domain, and then the modified z-transform applied to obtain this sum.

This is only one of numerous examples where the modified z-transform can be used to obtain the inverse Laplace more directly and with fewer calculations. As a rule, if the Laplace transform $C(s)$ has factorable terms in the form of e^{sTs} or $1/(1-e^{sTs})$ then it is advisable to inspect $C(s)$ further to see if the modified z-transform can yield quicker results.

It is hoped that this discussion serves to point out the general mathematical nature of the modified z-transform and its useful application to fields other than the sampled-data control systems. (A sampled-data system is shown in Fig. 13.)

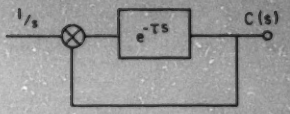


Fig. 16. Feedback control system with time-lag

Jack Sklansky (Radio Corporation of America, Princeton, N. J.): It is of interest to note that equation 15 yields the same answer for the output between sampling instants as that obtained by the method presented in a paper by the discusser and Dr. Ragazzini (see reference 4 of the paper), provided that one replaces the variable m by $(t-nT)/T$. The notation of the paper under discussion has the advantage of exhibiting more explicitly the variable m (making it convenient, for instance, to differentiate with respect to m); however, the labor involved in evaluating the residues in equation 15 is equivalent to that involved in evaluating the partial fractions of reference 4 of the paper. The latter paper has the advantage of providing a physical picture (by means of a block diagram corresponding to the partial fraction expansion), so that synthesis procedures for improving the intersample response are more readily brought to mind. The physical picture referred to is Fig. 4 in reference 4 of the paper, which is reproduced in this discussion as Fig. 14.

The root loci shown in Fig. 9 can be quite useful, but the discusser feels that the variables $\omega_0 T$ and η, which were used as parameters in plotting the loci, could with advantage have been replaced by the more convenient variables $\omega_0 T/2\eta$ and $1/2\omega_0\eta T$. The variables $\omega_0 T$ and η are the normalized natural frequency and the damping ratio respectively, in continuous systems (the normalization is with respect of the sampling frequency), but have only a vague analogous significance in sampled-data systems; on the other hand the variables $\omega_0 T/2\eta$ and $1/2\omega_0\eta T$ are the velocity constant and the normalized plant time constant, respectively (the normalization here is with respect to the sampling period), which have physical significance for all values of the sampling period, $T=0$ included.

The constant overshoot loci of Fig 10 are quite valuable, in the discusser's opinion. It would be of interest to plot these curves on the sT-plane by performing the transformation

$$sT=\log_e z \quad (95)$$

The discusser suspects that the resultant

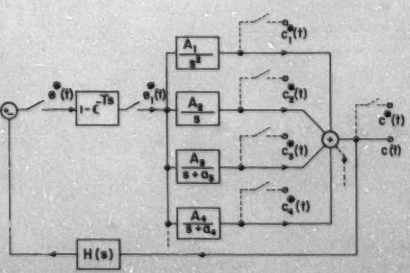


Fig. 14. Block diagram for determining the intersample response of an error-sampled feedback system (from reference 4 of the paper)

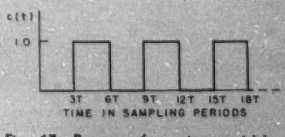


Fig. 17. Response for system containing time-lag

curves would be quite close to the constant-overshoot loci for the continuous case, which are straight lines through the origin of the sT-plane. The values of the real parts, in particular, of the sampled-data and the continuous loci should match more closely than those of the imaginary parts, because when the imaginary parts of the sampled-data loci in the sT-plane approach π, the sampling process swamps out the system dynamics as far as real frequency is concerned.

It might be well to point out that the symbol M_p, often used in the literature to represent the maximum frequency response ratio, in this paper represents the time domain overshoot. The symbol for maximum frequency response ratio which is used by the author in the subject paper is M_{max}.

Eliahu I. Jury: Mr. Farmanfarma notes that the modified z-transform may be successfully applied to obtain the inverse Laplace transform of certain functions. It is important to emphasize, as indicated by the discusser, that the application of the modified z-transform is especially advantageous in obtaining the inverse transform of functions which are rational in both e^{Ts} and s.

The example introduced by the discusser can be physically viewed as a sampled-data system as shown in Fig. 15. The modified z-transform of the output can be obtained as follows:

$$C^*(z,m)=\frac{1}{1-e^{-aT}}Z_m\frac{1}{s+a}=\frac{z}{z-1}\left(\frac{e^{-amT}}{z-e^{-aT}}\right) \tag{96}$$

Equation 96 is noticed to be exactly equivalent to equation 89 of the discusser. The continuous response may be obtained by finding the inverse of equation 96.

The response $C(t)$ may be obtained by using the inverse equation; however, the division method sometimes yields quicker results.

As an illustrative example, if the system in Fig. 16 is to be analyzed for a step input response, the modified z-transform of the output is given by

$$C^*(z,m)=\frac{e^{-m\tau}}{1+e^{-\tau s}}Z_m\frac{1}{s}=\frac{e^{-m\tau}}{1+e^{-\tau s}}\frac{1}{s-1} \tag{97}$$

If $\tau=3T$, equation 97 becomes

$$C^*(z,m)=\frac{1}{z^3+1}\frac{1}{z-1} \tag{98}$$

The actual time response can be easily obtained by the division method as plotted in Fig. 17. In this simple example the modified z-transform of the output is independent of m. However, if m appears in the modified z-transform of the output, still the division method (see reference 8 of the paper) may be applied.

As mentioned in the paper, there exist several methods of analysis which yield the output between the sampling instants. The connecting equation between the notation of the output expressed in this discussion and the time, as mentioned in the text, is the following equation

$$t=(n-1+m)T,\ 0\leq m\leq 1 \tag{99}$$

Dr. Sklansky has correctly correlated the notation m of the paper to the notations of his and Dr. Ragazzini's interesting paper. Equation 15 relates the output response quantities to the poles and zeros of the system function. This fact can be utilized advantageously for synthesis procedures, especially when digital compensators are utilized.[1] For instance, the integrand in equation 15 which might include the system function of the compensator, can be manipulated as much as possible to yield a prescribed response. Thus the compensator transfer function in the form of z-transform can be determined.

Tables II and III as well as Fig. 9 can still be utilized if the new convenient variables suggested by Dr. Sklansky are applied.

There is need for extending the existing methods of analyzing the actual response of sampled-data control systems so that these methods may be applied systematically to synthesis. It is hoped that the ideas embodied in this paper will initiate useful procedures to be applied in complementing the z-transform method.

REFERENCE

1. DISCRETE COMPENSATION OF SAMPLED-DATA AND CONTINUOUS CONTROL SYSTEMS, E. I. Jury, W. Schroeder. *Report No. 154*, Electronics Research Laboratory, University of California, Berkeley, Calif., series 60, 1955.

Factors in the Design of Digital Controllers for Sampled-Data Feedback Systems

JOHN E. BERTRAM
STUDENT MEMBER AIEE

Synopsis: This report is an extension of the design procedure proposed by Bergen and Ragazzini[1] for error-sampled control systems. Restrictions on the specification of the over-all transmission $K^*(z)$ caused by steady-state error considerations and the inability to cancel poles perfectly and zeros of the plant on or outside the unit circle are discussed. A design procedure which includes these restrictions is presented.

In addition, it is shown that where the input may take different forms it is desirable to have $K^*(z)$ a ratio of polynomials rather than a finite polynomial in z^{-1}. Several examples utilizing the specified design procedure are given.

SEVERAL recent papers[1-3] have been concerned with the utilization of a digital-controller or data-processing-unit sampled feedback-control system. One design procedure described by R. H. Barker[2,3] consists of adjusting the parameters of the data-processing unit in order to obtain a desired form of characteristic equation. In general, these characteristic equations are of two forms, first:

$$Z^k=0 \tag{1}$$

in which case it can be shown that the recovery following any transient will be complete k sampling periods after the end of the transient and, second:

$$(Z-\alpha)^k=0 \tag{2}$$

The latter form of characteristic equation can reduce overshoots and the response to random noise but at the expense of increased time of recovery following a transient. Either of the two forms of characteristic equation can be obtained with a data-processing unit, whose pulsed transfer function $D^*(z)$ has a constant in the numerator and polynomial in z^{-1} in the denominator. While the choice of the desired form of characteristic equation is motivated from time domain considerations, the technique certainly does not lead to direct time domain synthesis, since only the poles and not the zeros of the over-all transmission function are specified.

Bergen and Ragazzini,[1] however, present a method for direct time domain synthesis which utilizes a data-processing unit in systems similar to that of Fig. 1. By their procedure for a particular input, the program of the data-processing unit is so designed that the outputs at sampling instants are as specified. To obtain such complete control over the transient behavior of the system requires specification of both the zeros and poles of the over-all transmission function $K^*(z)$. To affect this design it is necessary that the

Paper 56-209, recommended by the AIEE Feedback Control Systems Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Winter General Meeting, New York, N. Y., January 30-February 3, 1956. Manuscript submitted October 17, 1955; made available for printing December 15, 1955.

JOHN E. BERTRAM is at Columbia University, New York, N. Y.

This research was supported in part by the U. S. Air Force under *Contract No. AF 18(600)-477* monitored by the Office of Scientific Research, Air Research and Development Command.

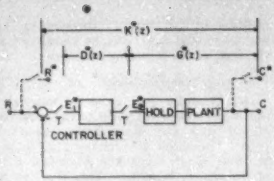


Fig. 1. Typical error-sampled feedback control system; T is the sampling period

pulsed transfer function of the data-processing unit be a ratio of polynomials in z^{-1}. Since every term in z^{-1} means one additional storage stage in the data-processing unit, the complexity of the controller for this system is about twice that for the Barker system which only adjusts the parameters of the characteristic equation.

To illustrate the design problems consider the system of Fig. 1. If the specified over-all transmission is $K^*_s(z)$, then it can easily be shown that the pulsed transfer function of the controller $D^*(z)$ must be

$$D^*(z)=\frac{1}{G^*(z)}\,\frac{K^*_s(z)}{1-K^*_s(z)} \tag{3}$$

For polynomial-type inputs (step, ramp, acceleration), Bergen and Ragazzini have tabulated over-all transmission prototypes which have finite settling time. However, the specification of $K^*_s(z)$ is not without restrictions.

In the present paper, these restrictions are reviewed and extended to the design of systems which do not necessarily have finite settling times. In addition, the problem of determining $K^*_s(z)$ when more than one type of input is expected is discussed.

Practical Limitations of Pole-Zero Cancellations

Time domain synthesis of error sampled control systems is achieved by canceling out the undesired poles and zeros of the pulsed transfer function $G^*(z)$ and replacing them with the desired poles and zeros in the compensated function $D^*(z)G^*(z)$. Unfortunately, perfect cancellation is impossible in any practical system. It has previously been shown[3] that any attempt to cancel a zero of $G^*(z)$ in the z-plane outside the unit circle leads to a system which is unstable in practice. By similar arguments it can be shown that the same is true of poles of $G^*(z)$. Actually, this limitation should extend to poles and zeros on, and even inside, but close to the unit circle.

To illustrate the effect of imperfect cancellation on or outside the unit circle in the z-plane, consider the pulsed transfer function of the plant

$$G^*(z)=\frac{(z-a_o)_G\prod_i(z-a_i)_G}{(z-b_o)_G\prod_i(z-b_i)_G} \tag{4}$$

where a_o is a zero and b_o a pole outside or on the unit circle, while the a_i's and b_i's are zeros and poles, respectively, inside the unit circle. For a specified over-all transmission function $K^*_s(z)$ the designer is led to a computer program with the pulsed transfer function of the form

$$D^*(z)=\left(\frac{(z-b_o)_D\prod_i(z-b_i)_D}{(z-a_o)_D\prod_i(z-a_i)_D}\right)\times\left(\frac{K^*_s(z)}{1-K^*_s(z)}\right) \tag{5}$$

where $K^*_s(z)$ and $[1-K^*_s(z)]$ contain the desired zeros and poles of the compensated function $D^*(z)G^*(z)$, if cancellation is perfect. Substituting equations 1 and 2 into the expression for the over-all transmission, $K^*(z)$, where

$$K^*(z)=\frac{D^*(z)G^*(z)}{1+D^*(z)G^*(z)} \tag{6}$$

it is found that

$$K^*(z)=\frac{(z-a_o)_G(z-b_o)_D\left[\prod_i(z-a_i)_G\prod_i(z-b_i)_D\right]K^*_s(z)}{\left[(z-a_o)_D(z-b_o)_G\left[\prod_i(z-b_i)_G\prod_i(z-a_i)_D\right]+\left\{(z-a_o)_G(z-b_o)_D\left[\prod_i(z-a_i)_G\prod_i(z-b_i)_D\right]-(z-a_o)_D(z-b_o)_G\left[\prod_i(z-a_i)_D\times\prod_i(z-b_i)_D\right]\right\}K^*_s(z)\right]} \tag{7}$$

If the cancellation of the plant poles and zeros and the digital controller zeros and poles were perfect, then equation 4 reduces after canceling the common terms to simply $K^*(z)=K^*_s(z)$, the specified transmission. However, this system has been achieved by the cancellation of two roots, a_o and b_o, of the characteristic equation which are on or outside the unit circle. That is, if $a_{oD}=a_{oG}$ and $b_{oD}=b_{oG}$ then the terms $(z-a_o)$ and $(z-b_o)$ may be factored out of the denominator and are cancelled by similar terms in the numerator. Now, assuming that the zeros and poles of the digital stabilizer remain as set, consider the effect of a shift of Δa_{oG} or Δb_{oG} in the zero or pole of the pulsed transfer function of the plant which are on or outside the unit circle. Since the characteristic equation (the denominator of $K^*(z)$) is simply a polynomial in z^{-1}, then the loci of the zeros as a function of any parameter is continuous. Therefore, for a parameter change of sufficiently small magnitude the zero of the characteristic equation is still outside the unit circle and the only possible way in which the system could remain stable is if a zero in the numerator shifted in unison with that in the denominator thereby canceling it. But, the numerator and denominator are different polynomials in z which, at most, can intersect at only a finite number of points. Therefore, for all practical purposes the system is unstable because the slightest parameter shift results in the characteristic equation having a zero outside the unit circle in the z-plane.

The same conclusions may be obtained by a somewhat different argument which proves useful in the case of poles and zeros inside but close to the unit circle. From equation 7 it is noted that the coefficient of $K^*_s(z)$ in the denominator is zero when the cancellation is perfect. For sufficiently small variations of the plant or controller parameters this term remains essentially zero. That is,

$$(z-a_o)_G(z-b_o)_D\left[\prod_i(z-a_i)_G\prod_i(z-b_i)_G\right]\approx(z-a_o)_D(z-b_o)_G\left[\prod_i(z-a_i)_D\prod_i(z-b_i)_D\right] \tag{8}$$

and equation 4 reduces to:

$$K^*(z)\approx\frac{(z-a_o)_G(z-b_o)_D\left[\prod_i(z-a_i)_G\times\prod_i(z-b_i)_D\right]K_s^*(z)}{(z-a_c)_D(z-b_o)_G\left[\prod_i(z-a_i)_D\times\prod_i(z-b_i)_G\right]} \tag{9}$$

From this expression it is readily apparent that all zeros and poles of the plant, including those outside or on the unit circle, appear in the characteristic equation. While the identical expression is to be found in the numerator, it is noted that each corresponding factor has a different subscript and, therefore, for perfect cancellation requires that the controller and plant singularities shift in unison which is not possible in practice.

Thus, for zeros and poles outside the unit circle, any continuous parameter variation causes the system to pass at least through a region of instability. For zeros and poles on the unit circle, the parameter variations have the possibility of moving them inside as well as outside the unit circle. However, the designer can hardly expect such good fortune, and actually he should consider carefully the

Table I. Zero Steady-State Error

Input	Pulsed Transform	Factor of $1-K^*(z)$ for Zero Steady-State Error
Step of position, $\mu(t)$	$\frac{1}{1-z^{-1}}$	$(1-z^{-1})$
Ramp or velocity, t	$\frac{Tz^{-1}}{(1-z^{-1})^2}$	$(1-z^{-1})^2$
Acceleration, t^2	$\frac{T^2z^{-1}(1+z^{-1})}{2(1-z^{-1})^3}$	$(1-z^{-1})^3$

cancellation of poles and zeros which are inside but close to the unit circle. From equation 9 it can be seen that to a first approximation the system is unstable if any of the zeros of the denominator term $i(z-a_i)_D$ should go outside the unit circle. Therefore, as a guide in design, the maximum allowable parameter variation for zeros and poles inside the unit circle is given by

$$\begin{aligned} |b_i+\Delta b_i|_G &< 1 \\ |a_i+\Delta a_i|_D &< 1 \end{aligned} \tag{10}$$

Generally, the digital controller can be constructed sufficiently well that Δa_{iD} will be held to zero so that the problem is primarily concerned with the stability of poles of the plant just inside the unit circle. However, poles and zeros of the plant on or outside the unit circle are equally bad, for here complete cancellation is necessary if the system is to be stable.

By reference to equation 5, it is apparent that if the designer of the digital stabilizing unit is not to attempt the cancellation of poles and zeros of the pulsed transfer function of the plant on or outside the unit circle, then $K^*_s(z)$ must contain all such zeros and $[1-K^*_s(z)]$ must contain all such poles. With these restrictions, equation 5 becomes

$$D^*(z)=\left(\frac{(z-b_o)_D\prod_i(z-b_i)_D}{(z-a_o)_D\prod_i(z-a_i)_D}\right)\times\left(\frac{(z-a_o)_DA^*(z)}{(z-b_o)_DB^*(z)}\right) \tag{11}$$

where

$$(z-a_o)_DA^*(z)=K^*_s(z) \tag{12A}$$

$$(z-b_o)_DB^*(z)=1-K^*_s(z) \tag{12B}$$

and $D^*(z)G^*(z)$ is now given by

$$D^*(z)G^*(z)=\left[\frac{\prod_i(z-b_i)_DA^*(z)}{\prod_i(z-a_i)_DB^*(z)}\right]\times\left[\frac{(z-a_o)_G\prod_i(z-a_i)_G}{(z-b_o)_G\prod_i(z-b_i)_G}\right] \tag{13}$$

If the cancellation of poles and zeros inside the unit circle is considered perfect (as previously stated, the designer must be careful if the zeros of the plant are close to the unit circle) the over-all transmission reduces to simply

$$K^*(z)=(z-a_o)_GA^*(z) \tag{14}$$

as specified.

In summary, for a unity feedback error sampled control system, the following design restrictions must be considered.

1. Any zeros of the plant on or outside the unit circle in the z-plane must be contained in $K^*_s(z)$.
2. Any poles of the plant on or outside the unit circle in the z-plane must be contained in $[1-K^*_s(z)]$.
3. If the expected variation of any plant pole, $|\Delta b_i|$ is greater than or equal to $|1-|b_i||$ then this zero should be considered as falling outside the unit circle and appears as a factor of $K^*_s(z)$.

Steady-State Error Considerations

In order to obtain zero steady-state error in response to test input of the form t^n it is necessary to impose additional conditions on equation 12(A) for the term $[1-K^*_s(z)]$.

The sampled system error $E^*_1(z)$ for the error sampled system of Fig. 1 is given by

$$E_1^*(z)=R^*(z)[1-K^*(z)] \tag{15}$$

The expression for the sampled steady-state error following a test input is obtained from equation 15 by applying the final value theorem

$$\lim_{m\to\infty}E^*(mT)=\lim_{z^{-1}\to 1}(1-z^{-1})R^*(z)\times[1-K^*(z)] \tag{16}$$

For a step input where $R^*(z)=1/1-z^{-1}$, in order that the $\lim_{m\to\infty}E^*(mT)=0$, it is necessary that the term $[1-K^*(z)]$ contain as a factor $(1-z^{-1})$. In a similar fashion for a ramp input, $R^*(z)=Tz^{-1}/(1-z^{-1})^2$, it is necessary that the term $[1-K^*(z)]$ contain as a factor $(1-z^{-1})^2$ in order that the steady-state error be zero. In a similar manner the requirement for zero steady-state error to higher order test inputs may be determined. These results are summarized in Table I.

It follows directly that a system which follows an input t^n with zero steady-state error follows all inputs t^k where k is any integer less than n with zero steady-state error. Thus, a system which follows a ramp with zero steady-state error does the same for a step.

The appearance of these factors $(1-z^{-1})$, in $[1-K^*_s(z)]$ compares in the continuous case to the presence of integrations in the forward path. From equation 5 for $D^*(z)$ it is seen that any integrations in the plant-hold combination which add factors $(1-z^{-1})$ in the numerator cancel out the need of such action in the controller. For example, if the system is to follow a ramp input with zero steady-state error, then $[1-K^*(z)]$ must contain the factor $(1-z^{-1})^2$. However, if the plant contains two integrations then the numerator of $D^*(z)$ also contains the factor $(1-z^{-1})^2$ and the controller needs no further integrations. Bergen and Ragazzini point out that the controller should not be required to have more than one integration. This is a practical limitation which can sometimes be ignored if the input is truncated before the plant saturation level is reached.

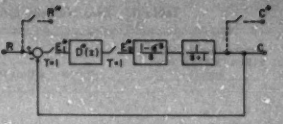


Fig. 2. A system in which the plant has no integrations but which follows a truncated ramp with zero steady-state error

To illustrate this point, consider the system of Fig. 2, which is to follow a ramp at R^* with zero steady-state error at sampling instants. The plant of this system has no integrations, so the digital stabilizer must contain two, in order that the steady-state error to a ramp input be zero. Bergen and Ragazzini have automatically ruled out this possibility, but consider the output of this controller. The z-transform of the pulse sequence out of the digital controller following the input t is

$$E^*_2(z)=\left(\frac{2}{0.632}\right)\times\left(\frac{z^{-1}-0.868z^{-2}+0.184z^{-3}}{(1-z^{-1})^3}\right) \tag{17}$$

From the inverse transform, it is seen that the first few terms of this pulse sequence are

$$E_2(mT)=0\delta(0)+3.16\delta(T)+3.58\delta(2T)+4.58\delta(3T)+5.58\delta(4T)+6.58\delta(5T)+\ldots \tag{18}$$

In this case, if the plant saturated for inputs of ± 50 units, for the present slope of input the output would follow accurately for 48 samples before saturation. If this were the maximum input slope and the function is always truncated in less than 48 sample intervals then the restriction on more than one integration in the data-processing unit could be ignored.

Table II. Zero Velocity Error System with Finite Settling Time

Parameters for Min $\Sigma[E(mT)]^2$ to a Step Input	$\Sigma[E(mT)]^2$ For a Unit Pulse	For a Unit Step	For a Unit Ramp
$b_0=0$	6	2	1
$b_1=1/2$	$3^1/_2$	$1^1/_2$	$1^1/_4$
$b_1=2/3,\ b_2=1/3$	$2^9/_{10}$	$1^1/_5$	$1^9/_{10}$
$b_1=3/4,\ b_2=1/2,\ b_3=1/4$	$2^9/_4$	$1^1/_4$	$1^7/_8$
$b_1=\frac{k}{1+k},\ b_2=\frac{k-1}{k+1}\cdots$ $b_k=\frac{1}{k+1}$	$\frac{2k^2+6k+6}{(1+k)^2}$	$\frac{2+k}{1+k}$	$\frac{2k^2+9k^2+13k+6}{6(1+k)^2}$

Table III. Zero Position Error System

Input	$E_1^*(z)$	Steady-State Value
Step, $u(t)$	$\frac{1}{1+c_1z^{-1}}$	0
Ramp, t	$\frac{Tz^{-1}}{(1-z^{-1})(1+c_1z^{-1})}$	$\frac{T}{1+c_1}$
Acceleration, t^2	$\frac{T^2}{2}\frac{z^{-1}(1-z^{-1})}{(1-z^{-1})^2(1+c_1z^{-1})}$	∞

However, each case must be considered separately to determine if the desired steady-state specifications can be met in practice.

General Design Equations

Equations 12(A) and 12(B) for $K^*(z)$ and $[1-k^*(z)]$ together with the restraints outlined in the preceding sections may be directly utilized for the design of error-sampled systems. Since $R^*(z)$ must simultaneously satisfy both equations 12(A) and 12(B), either or both of which are constrained in part, it is apparent that the designer does not always have complete freedom of choice in the response to a specified input.

To illustrate the problem consider the system of Fig. 1, which is to follow a ramp input applied at R with zero steady-state error at sampling instants. If the settling time is to be finite, then both $K^*(z)$ and $[1-K^*(z)]$ must be finite polynomials in z^{-1} and are written as

$$K^*(z)=(z-a_0)_DA^*(z)=a_1z^{-1}+a_2z^{-2}+\ldots+a_mz^{-m} \quad (12A)$$

and

$$1-K^*(z)=(z-b_0)_DB^*(z)=(1-z^{-1})^2(1+b_1z^{-1})+\ldots+b_{m-2}z^{-(m-2)} \quad (12B)$$

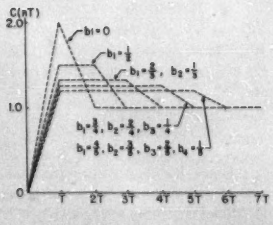


Fig. 3. The transient response to a unit step for zero velocity error system with finite settling time and minimum $\sum_{m=0}^{\infty}[E_1(mT)]^2$

Since $K^*(z)$ must simultaneously satisfy both equations 12(A) and 12(B), it is apparent that both expressions must be of the same power in z^{-1}. In the present case, $K^*(z)$ must contain at least two terms, since $[1-K^*(z)]$ requires the factor $(1-z^{-1})^2$. This assures that all zeros of the plant are inside the unit circle. Solution of this simple problem results in $a_1=2$ and $a_2=-1$, which is the minimum prototype response for a system subjected to a ramp input. However, if a step input is applied to this system there is a 100-per-cent overshoot at the first sample instant before settling down to the desired value. This is a rather severe transient and greatly limits the usefulness of such a system. A possible method of reducing the overshoot is to add additional terms to $K^*(z)$ which automatically adds the same number of terms to $[1-K^*(z)]$. One disadvantage of this solution is that for each additional term in $K^*(z)$ the digital controller transfer function is increased by two terms, which means the addition of two storage stages in the computer.

Moreover, if more terms are to be added, there remains the problem of setting the parameters a_i and b_i. Because of the mathematical ease, the optimum system has been chosen as that which has the minimum sum of squared errors at sampling instants. From the inversion theorem, this sum has been shown[2] to be

$$\sum_{m=0}^{\infty}[E_1(mT)]^2=\frac{1}{2\pi j}\int_{\Gamma}E_1^*(z)E_1^*(z^{-1})\frac{dz}{z} \quad (19)$$

where Γ is the unit circle in the z-plane. The evaluation of this integral simply reduces to finding the residues for all singularities within the unit circle. However, for a system designed for finite settling time, the only singularities are at the origin $z=0$. As a result, it is perhaps easier to consider the pulsed error transform

$$E^*(z)=R^*(z)[1-K^*(z)] \quad (20)$$

which is a rational finite polynomial in z^{-1}. The sum of the squared error, $\sum_{m=0}^{\infty}[E(mT)]^2$, is simply the sum of the squares of the coefficients of equation 20 as long as it is a finite polynomial in z^{-1}. Thus, for the present example with the minimum $K^*(z)=2z^{-1}-z^{-2}$, the sum of the squared errors at sampling instants is 2 for a step input, $R(z)=1/1-z^{-1}$. If one extra term is added to $[1-K^*(z)]$ in order to reduce the overshoot to a step input the pulsed error transform becomes

$$E^*(z)=z^{-1}-(1-b_1)z^{-1}-b_1z^{-2} \quad (21)$$

and the sum of the squared errors is simply

$$\sum_{m=0}^{\infty}[E(mT)]^2=2(1-b_1+b_1^2) \quad (22)$$

When this sum is minimized with respect to b_1 the result is $\sum_{m=0}^{\infty}[E(mT)]^2=1.5$ for $b_1=0.5$. Table II shows the result of adding additional terms. From this table it is noted that the error $\sum_{m=0}^{\infty}$

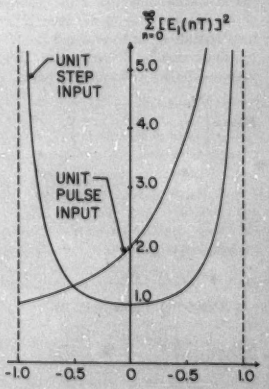


Fig. 4. A zero position error system, transient performance as a function of c_1

Table IV. Zero Velocity Error System

Input	$E_1^*(z)$	Steady-State Error
Unit step, (l)	$\frac{1-z^{-1}}{1+c_1z^{-1}+c_2z^{-2}}$	0
Unit ramp, t	$\frac{Tz^{-1}}{1+c_1z^{-1}+c_2z^{-2}}$	0
Unit acceleration, t^2	$\frac{Tz^{-1}}{(1-z^{-1})(1+c_1z^{-1}+c_2z^{-2})}$	$\frac{T^2}{1+c_1+c_2}$

$[E(mT)]^2 \to 1$ as $k \to \infty$ for a step input, $R^*(z) = 1/1-z^{-1}$. However, these same b_i's result in the $\sum_{m=0}^{\infty} [E(mT)]^2$ for a ramp input, $R^*(z) = z^{-1}/(1-z^{-1})^2$, constantly increasing with k. At the same time the sum of the squared errors to a unit pulse input is approaching 2 as $k \to \infty$. This latter sum gives an indication of the system performance in the presence of noise. It is also interesting to note that the b_i's for optimum performance to a step input in the least squares sense maintain the error or overshoot in this case at a constant value until the steady state is reached. Fig. 3 shows this characteristic for the system of Table II. The parameters b_i may also be set such that they minimize the sum of the squares of the error to a pulse input, but the trend and magnitudes as k increases is not very different from the case considered.

If additional restraints on $k^*(z)$ and $[1-K^*(z)]$ are imposed by the poles and zeros of the pulsed transfer function of the plant, a similar analysis may be performed to determine the coefficients for best performance in the sense of the minimum sum of the squared error at sampling instants. However, it must not be forgotten that this improvement has been derived from additional complexity in the digital controller.

Equal performance in the least squares sense can be obtained by scarificing finite settling time and adding a denominator to $K^*(z)$ and $[1-K^*(z)]$. Further, if the highest power of this denominator in z^{-1} is no greater than the highest power of the numerator, the number of terms in the pulsed transfer function of the digital computer is unchanged by this addition. This can be seen easily by referring to equation 5 for $D^*(z)$ and noting that $K^*(z)$ and $[1-K^*(z)]$ appear as a ratio so that their common denominators are cancelled.

With the addition of a denominator the general design equations are somewhat modified. To illustrate the general problem consider a system which is to follow an input t^ρ with zero steady-state error and where plant pulsed transfer function has k poles α_k, and i zeros β_i outside the unit circle. With these restraints the design relations become

$$K^*(z) \simeq \frac{\prod_i (1-\alpha_i z^{-1})(a_0+a_1z^{-1}+\ldots+a_{m-k}z^{-(m-k)})}{1+c_1z^{-1}+c_2z^{-2}+\ldots+c_mz^{-m}} \quad (23)$$

$$1-K^*(z) = \frac{\prod_i (1-\beta_i z^{-1})(1+b_1z^{-1}+\ldots+b_{m-l-\rho}z^{-(m-l-\rho)})(1-z^{-1})^\rho}{1+c_1z^{-1}+c_2z^{-2}+\ldots+c_mz^{-m}} \quad (24)$$

The simultaneous solution of these two relations lead to n linear algebraic equations with $[2n-(l+k+\rho)]$ unknowns. Therefore, $[n-(l+k+\rho)]$ parameters may be set arbitrarily from the purely mathematical considerations. However, the added denominator is the characteristic equation of the system, so that the c_i's are constrained such that all zeros of this expression in the z-plane are within the unit circle. For finite settling time all c_i's are set equal to zero. Another characteristic equation which is often used by British designers is of the form $(1-cz^{-1})^m$ where c is called the staleness factor.

To illustrate the advantages of finite values of c_i in the denominator of $K^*(z)$ and $[1-K^*(z)]$ consider a plant $G^*(z)$ which has no poles or zeros outside or on the unit circle except for one pole at $z=1$. This plant is to be incorporated first in a system which has zero displacement error in the steady state and second, in a system which has zero velocity error in the steady state.

Zero Displacement Error Systems

Since the plant contains at most a ingle integrator, its pulsed transfer function has no poles or zeros outside the unit circle except for the pole at $z=1$; the design equations for a minimum polynomial $D^*(z)$ are simply

$$K^*(z) = \frac{a_1z^{-1}}{1+c_1z^{-1}} \quad (25)$$

$$1-K^*(z) = \frac{1-z^{-1}}{1+c_1z^{-1}} \quad (26)$$

The combination and equation of the coefficients of z^{-1} leads to the condition that (a_1+c_1) must equal unity, and from stability considerations the magnitude of $|c_1|<1$. The z-transform of the system error $E_1^*(z)$ to a unit step, a unit ramp, and a unit acceleration, and the steady-state error to each is shown in Table III.

Here it is seen that the addition of the denominator affects the steady-state error in following a ramp. In the prototype design with $c=0$ the error in following a unit ramp is T. This value can be reduced if c_1 is a positive constant. This expression is analogous to the reciprocal of the velocity constant in a continuous servomechanism. As before, the sum of the squared errors at sampling instant will be considered as indicative of the transient performance.

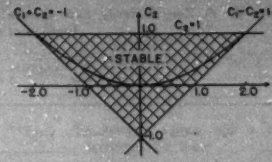


Fig. 5. The stability triangle for a system with a characteristic equation of the form $z^2+c_1z+c_2=0$

When a step input is applied this value can be obtained from

$$\sum_{m=0}^{\infty} [E(mT)]^2 = \left(\frac{1}{2\pi J}\int_\Gamma \frac{1}{1+c_1z^{-1}}\right) \times \left(\frac{1}{1+c_1z}\frac{dz}{z}\right) \quad (27)$$

where Γ is again the unit circle in the z-plane. Evaluation of this expression yields

$$\sum_{m=0}^{\infty} [E(mT)]^2 = \frac{1}{1-c_1^2} \quad (28)$$

where for stability c_1 is confined to the range $-1<c_1<+1$. In this case the sum of the squared pulsed errors is a minimum for $c_1=0$. Fig. 4 shows graphically the results over the range of possible c_1. Note that while the sum is increased by a third for $c_1=0.5$, the steady-state error to a ramp input is reduced by the same factor. In some design problems this additional flexibility with no extra expenditure of components in the digital controller may be desirable.

Since the response to a single pulse may be considered characteristic of the noise response of the system, the sum of the squared error for an impulse applied at $t=0$ is also plotted on Fig. 4. This result is obtained in the same way as the previous sum and may be expressed as

$$\sum_{m=0}^{\infty} [E(mT)]^2 = \frac{2}{1-c_1} \quad (29)$$

It is apparent in this case that low steady-state error to a ramp input and response to pulse-type disturbances is incompatible, and a final decision as to the choice of c_1 will depend on the relative importance attached to the various transients and the steady-state error.

Table V. A Zero Velocity Error System: Transient Performance to a Step Input as a Function of c_1 and c_2

$$\sum_{m=0}^{\infty} [E(mT)]^2 \frac{2}{1-c_2^2+c_1(c_2-1)}$$

c_2 \ c_1	−1	−0.8	−0.6	−0.4	−0.2	0	0.2	0.4	0.6	0.8	1
−2											
−1.8										2.78*	
−1.6									1.56*	2.94	
−1.4								1.19*	1.67	3.12	
−1.2							1.04*	1.35	1.78	3.33	
−1.0						1*	1.14	1.39	1.92	3.57	
−0.8					1.04*	1.11	1.25	1.52	2.08	3.84	
−0.6				1.19*	1.19	1.25	1.39	1.67	2.28	4.17	
−0.4			1.56*	1.43	1.39	1.43	1.56	1.85	2.50	4.55	
−0.2		2.78	2.08	1.79	1.67	1.67	1.79	2.08	2.78	5.00	
0	∞	5.55	3.12	2.38	2.08	2	2.08	2.38	3.12	5.55	
0.2		∞	6.25	3.58	2.78	2.5	2.5	2.78	3.56	6.25	
0.4			∞	7.16	4.18	3.34	3.13	3.34	4.16	7.14	
0.6				∞	8.35	5.0	4.17	4.16	5.00	8.34	
0.8					∞	10.0	6.25	5.55	6.25	10.00	
1.0						∞	12.50	8.34	8.34	12.50	
1.2							∞	16.70	12.50	16.70	
1.4								∞	16.70	25.00	
1.6									∞	50.00	
1.8										∞	
2											

Zero Velocity Error Systems

A more interesting problem which illustrates to an even greater degree the advantages of making $K^*(z)$ a ratio of polynomials is the design of a system, containing the plant of the previous example, to follow a ramp or velocity with zero steady-state error. With these restraints the minimum expressions for $K^*(z)$ and $[1-K^*(z)]$ are

$$K^*(z) = \frac{a_1 z^{-1} + a_2 z^{-2}}{1 + c_1 z^{-1} + c_2 z^{-2}} \tag{30}$$

and

$$1 - K^*(z) = \frac{(1-z^{-1})^2}{1 + c_1 z^{-1} + c_2 z^{-2}} \tag{31}$$

The solution of these expressions simultaneously and the equation of the coefficients of like powers of z^{-1} gives the expressions for a_1 and a_2.

$$\begin{aligned} a_1 &= 2 + c_1 \\ a_2 &= -1 + c_2 \end{aligned} \tag{32}$$

The restrictions on c_1 and c_2 for stability are not quite so obvious as in the previous case. However, upon application of the Routh-Hurwitz criterion it can be shown that the following must be satisfied in order that the system be stable.

$$\begin{aligned} &c_2 < 1 \\ &1 + c_1 + c_2 > 0 \\ &1 - c_1 + c_2 > 0 \end{aligned} \tag{33}$$

These three inequalities may be used to construct the triangle shown in Fig. 5 in the c_1 and c_2 parameter plane. It is necessary that c_1 and c_2 lie inside the triangle in order that the system be stable.

A great deal of interesting information can be derived from the stability triangle. For example, the parabola passing through the origin and two vertices of the triangle is the locus of all points for which the characteristic equation has equal roots. This corresponds to the form of characteristic equation of equation 2 where $k=2$ while the origin corresponds to the form of equation 1. Values of c_1 and c_2 above the parabola lead to a characteristic equation with two complex conjugate roots, while values below result in two real roots.

The pulsed system error transform and the steady-state error to a unit step, a unit ramp, and a unit acceleration are shown in Table IV. For the finite settling time case where $c_1=c_2=0$ the steady-state error in following an acceleration is T^2. This error can be reduced by the proper choice of c_1 and c_2. This is analogous to improving the acceleration constant in a continuous servomechanism.

Again, by utilization of the sum of the squared errors at sampling instants as a measure of the transient performance one finds that the result is a function of both c_1 and c_2. Since c_1 and c_2 are constrained to values within the stability triangle of Fig. 5 the $\sum_{m=0}^{\infty} [E(mT)]^2$ only has meaning within this range of parameters. It will be found that the minimum sum to a step input will occur on the boundary of the triangle. Obviously this solution is not acceptable and a true minimum does not exist.

At this point another advantage to making $K^*(z)$ a ratio of polynomials is found. From the minimum prototype design for finite settling time, it is known that while the transient response to a ramp is quite satisfactory, the response to a step input overshoots by 100 per cent before settling down to the steady state. In many systems this would be intolerable and as previously shown may be reduced by adding additional terms to $K^*(z)$. However, by the addition of a denominator the parameters c_1 and c_2 may be so adjusted that better performance may be obtained and with no additional components or stages in the digital controller.

The sum of the system squared error

Table VI. A Zero Velocity Error System: Transient Performance to Unit Ramp as a Function of c_1 and c_2

$$\sum_{m=0}^{\infty} [E(mT)]^2 \frac{(1+c_2)}{(1-c_2)} \times \frac{1}{[(c+c_2)^2 - c_1^2)]}$$

c_2 \ c_1	−1	−0.8	−0.6	−0.4	−0.2	0	0.2	0.4	0.6	0.8	1
−2											∞
−1.8										∞	∞
−1.6									∞	13.2	∞
−1.4								∞	6.67	7.00	∞
−1.2							∞	4.50	3.57	5.00	∞
−1.0						∞	3.41	2.43	2.86	4.00	∞
−0.8					∞	2.78	1.88	1.77	2.08	3.57	∞
−0.6				∞	2.38	1.56	1.39	1.46	1.82	3.13	∞
−0.4			∞	2.14	1.39	1.19	1.17	1.30	1.67	2.92	∞
−0.2		∞	2.08	1.34	1.11	1.04	1.07	1.22	1.59	2.82	∞
0	∞	2.78	1.56	1.19	1.04	1	1.04	1.19	1.56	2.78	∞
0.2		∞	2.08	1.34	1.11	1.04	1.07	1.22	1.59	2.82	∞
0.4			∞	2.14	1.39	1.19	1.17	1.30	1.69	2.96	∞
0.6				∞	2.38	1.56	1.39	1.46	1.82	3.13	∞
0.8					∞	2.78	1.88	1.77	2.08	3.57	∞
1.0						∞	3.41	2.43	2.56	4.00	∞
1.2							∞	4.50	3.57	5.00	∞
1.4								∞	6.67	7.00	∞
1.6									∞	13.2	∞
1.8										∞	∞
2											∞

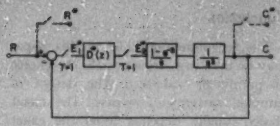


Fig. 6. An example system with two integrations in the plant

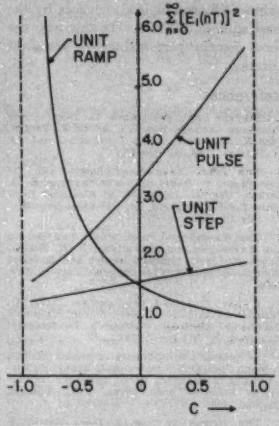


Fig. 7. Transient performance of system with two integrations as a function of c

pulse train for a step input is given by

$$\sum_{m=0}^{\infty}[E(mT)]^2=\frac{2}{1-c_1^2+c_1(c_1-1)} \tag{34}$$

while for a ramp input it is

$$\sum_{m=0}^{\infty}[E(mT)]^2=\frac{1}{(1+c_1)^2-c_1^2} \tag{35}$$

These results may be obtained directly by integrating equation 24 after substituting the appropriate pulsed error transform.

Thus, if it is equally likely that this system is to receive either step or ramp inputs Tables V and VI may be used to effect a solution which is a compromise in the performance of each. For example, if c_1 is chosen equal to -0.5 and c_2 equal to zero, the sum of the squared error to both a step and a ramp is 1.33. Of course, better results for a step can be achieved but at the expense of worse performance in the presence of a ramp.

Therefore, it would seem that the best system in the least squares sense for both step and ramp inputs would have a characteristic equation of the form $z(z-0.5)=0$. This conclusion, in preference to the traditional forms of characteristic equations,[2] is further borne out by the sum of the squared errors to a unit pulse input. With a characteristic equation designed for finite settling time, $z^2=0$, the sum of the squares of the error is

$$\sum_{m=0}^{\infty}[E(mT)]^2=6 \tag{36}$$

While the characteristic equation of the form $(z-\alpha)^2$, where $\alpha=c_1/2$ results in this sum being

$$\sum_{m=0}^{\infty}[E(mT)]^2=6+2c_1 \tag{37}$$

This is a minimum within the stability triangle for $c_1\approx -2$, but the sum of the squares of the error for both a step and a ramp input are quite high in this vicinity of c_1. However, for the characteristic equation $z(z+c_1)$

$$\sum_{m=0}^{\infty}[E(mT)]^2=\frac{6+2c_1}{1-c_1} \tag{38}$$

This expression has a minimum value of approximately 2 within the stability triangle for c_1 in the region of -1.

Again the final selection of the characteristic equation depends upon the expected inputs and their relative importance.

An Example

To illustrate how the latter form of characteristic equation might aid the design of practical systems, consider Fig. 6. Here the plant is a double integration which is inherently unstable. The pulsed transform of the plant-hold combination for a sampling rate of 1 second is $z^{-1}(1+z^{-1})/(1-z^{-1})^2$. Thus, the plant has a zero and two poles on the unit circle. $K_s^*(z)$ must contain the factor $(1-z^{-1})^2$. As a result of the latter, the system will follow a ramp with zero steady-state error. If a characteristic equation of the form $z^2(z+c)$ is used, then

$$K_s^*(z)=\frac{(1+z^{-1})(a_1z^{-1}+a_2z^{-2})}{(1+cz^{-1})} \tag{39}$$

and

$$1-K_s^*(z)=\frac{(1-z^{-1})^2(1+b_1z^{-1})}{(1+cz^{-1})} \tag{40}$$

are the resulting design equations. If fewer parameters than a_1, a_2, and b_1 were used it would be found that $|c|$ is required to be greater than 1, which is impossible if the system is to be stable.

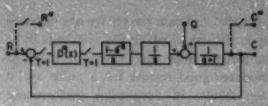


Fig. 8. An example system with a disturbance-type input

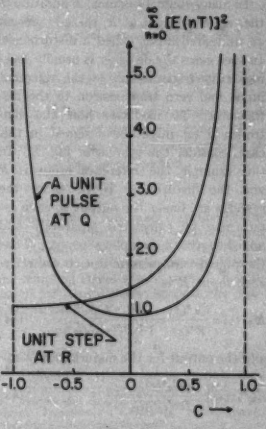


Fig. 9. Transient performance of system with disturbance-type input

By substitution of equation 39 into equation 40 and solving for b_1 in terms of c, it is found that

$$b_1=\frac{3+c}{4} \tag{41}$$

Thus, the z-transform of the system error becomes

$$E_s^*(s)=R^*(s)\left[\frac{(1-z^{-1})^2\left(1+\frac{3+c}{4}z^{-1}\right)}{(1+cz^{-1})}\right] \tag{42}$$

The sum of the squares of this error at sampling instants as a function of c for a unit pulse, a unit step, and a unit ramp are, respectively,

$$\sum_{m=0}^{\infty}[E(mT)]^2=\frac{6c^2+36c+54}{16}\text{(unit pulse)},\quad \frac{6c+26}{16}\text{(unit step)},\quad \frac{7c+25}{16(1+c)}\text{(unit ramp)} \tag{43}$$

These functions are plotted in Fig. 7 for

$-1<c<1$. Again the choice of c depends upon the use to which the system is to be subjected, since the value of c which produces good results for a pulse and a step results in very poor performance when a ramp is applied.

Second Example

In many control systems, in addition to the reference input at R, there is present an undesired input called a disturbance. In such cases the designer is usually seeking unity transmission to the reference input and zero transmission to the disturbance. To illustrate how the controller $D^*(z)$ might be designed in this case, consider the system of Fig. 8. In this example the reference input is R and the disturbance is Q. Since the systems are linear the output to each input may be calculated separately and added to give the complete output. Thus the pulsed transmission due to the reference input R is

$$K_R^*(z)=\frac{C_R^*(z)}{R^*(z)}=\frac{D^*(z)G_1G_2^*(z)}{1+D^*(z)G_1G_2^*(z)} \qquad (44)$$

and the output for the disturbance Q is

$$C_Q^*(z)=\frac{QG_2^*(z)}{1+D^*(z)G_1G_2^*(z)} \qquad (45)$$

In the design of such a system, first consider the transmission of the reference input R. The pulsed transfer function of the plant is

$$G_1G_2^*(z)=\frac{0.368z^{-1}(1+0.72z^{-1})}{(1-z^{-1})(1-0.368z^{-1})} \qquad (46)$$

If the system is to respond to a step input at R with zero steady-state error, then with the minimum complexity $D^*(z)$ for this constraint

$$K_R^*(z)=\frac{az^{-1}}{1+cz^{-1}} \qquad (47A)$$

$$1-K_R^*(z)=\frac{1-z^{-1}}{1+cz^{-1}} \qquad (47B)$$

By solution of these relations simultaneously it is found that $a=1+c$.

This system has previously been investigated and the sum of the error squared at sampling instants for a step input $R^*(z)=1/(1-z^{-1})$ is

$$\sum_{m=0}^{\infty}[E_R(mT)]^2=\frac{1}{1-c^2} \qquad (48)$$

Now the output $C^*_Q(z)$ due to the disturbance, is entirely an error in desired output. By combination of equations 44 and 45 it is possible to express the transform of the undesired output as

$$C_Q^*(z)=E_Q^*(z)=QG_2^*(z)[(1-K_R^*(z)] \qquad (49)$$

If the disturbance Q is noise in the form of impulses one may again calculate the sum of the squared errors at sampling instants resulting from the application of pulse at a sampling instant. This sum is

$$\sum_{m=0}^{\infty}[E_Q(mT)]^2=\frac{1.463}{(1-c)(1+0.368c)} \qquad (50)$$

In Fig. 9 the sums of equations 48 and 50 are plotted as functions of c. It is seen that values of c which will be chosen lie in the vicinity of -0.5.

Conclusion

The techniques outlined in this paper may be used to narrow greatly the choice of parameter values in the design of sampled systems. However, the exact choice usually depends upon the exact application. As an aid to the final choice since inversion is so simple in the z-plane, it is possible actually to calculate the transient response for several parameter values indicated as likely choices by the least squares criterion. In this way the time domain performance is the final criterion.

References

1. Sampled-Data Processing Techniques for Feedback Control Systems, Arthur R. Bergen, John R. Ragazzini. *AIEE Transactions*, vol. 73, pt. II, Nov. 1954, pp. 236–47.

2. The Pulse Transfer Function and Its Application to Sampling Servo Systems, R. H. Barker. *Journal*, Institution of Electrical Engineers, London, England, pt. IV, Monograph No. 43, July 15, 1952.

3. The Theory of Pulse-Monitored Servos and Their Use for Prediction, R. H. Barker. *Report No. 1046*, Signals Research and Development Establishment, Christchurch, England, Nov. 1950.

4. Linear Filtering of Sampled Data, G. Franklin. *Technical Report T-5/B*, Columbia University Electronics Research Laboratories, New York, N. Y., Dec. 1, 1954.

5. Automatic and Manual Control (book), C. Smith, Holt, D. F. Lawden, A. E. Bailey. "Characteristics of Sampling Servo Systems," Butterworths Scientific Publications, Ltd., London, England, 1952.

6. The Design of Sampled-Data Feedback Systems, Gladwyn V. Lago, John G. Truxal. *AIEE Transactions*, vol. 73, pt. II, Nov. 1954, pp. 247–53.

7. Fundamental Theory of Servomechanisms (book), L. A. MacColl. D. Van Nostrand Company, Inc., New York, N. Y., 1945.

8. A New Approach to the Design of Pulse-Monitored Servo Systems, A. Porter, F. Stoneman. *Journal*, Institution of Electrical Engineers London, England, vol. 97, pt. II, 1950, pp. 597–610.

Discussion

Eliahu I. Jury and **W. Schroeder** (University of California, Berkeley, Calif.): The main contribution of this paper is its indication that if a system is to follow both step and ramp inputs, a compromise between settling time and overshoot at the sampling instants should be made. This can be accomplished, as the paper clearly points out, by adding zeros to the over-all transfer function or by forcing the roots of the characteristic equation not to coincide at the origin.

Both this paper and that of Bergen and Ragazzini (see reference 1 of the paper) are mainly concerned with the response at the sampling instants. However, the actual response of the plant is continuous and therefore the response between the sampling instants is of considerable importance and should not be overlooked. In this connection, the author's statement in referring to the control at the sampling instants as "such complete control over the transient behavior of the system" should be modified. Control of the response at the sampling instants is indeed not complete control of the actual response.

The discussers of a recent paper[1] have shown that to obtain better control over the actual response, only the poles of the plant transfer function may be cancelled by the processing unit. Any attempt to cancel a zero only results in the addition of an undesirable pole, no matter how precisely the pole is fixed. This stems from the fact that the zeros of the original plant in the z-domain are not fixed but are dependent on where in time the output is specified. This is indicated clearly in a paper on the modified z-transform.[2] Any attempt to cancel a zero by adding a pole, as suggested by the author, can only result in a large settling time for the continuous response.

To illustrate the uncertainty of the use of the z-transform throughout this paper as well as in reference 3 of the paper, reconsider the first example chosen by Bergen and Ragazzini and repeated in Fig. 10.

The original and compensated responses are shown in Fig. 11. Curve b is obtained by Bergen and Ragazzini with a compensator whose transfer function is (equation 37 of reference 3 of the paper)

$$G_c^*(z)=\frac{(z-0.368)z^2}{(0.528z+0.104)(z-1)(z+1)} \qquad (51)$$

which forces the error to be zero at the sampling instants with no regard for the complete response. As seen in Fig. 11 the continuous response of this system has an overshoot of 20 per cent and a large settling time. In comparison, if the method presented in reference 1 of this discussion is extended to a system with arbitrary delays, the compensator and response may be de-

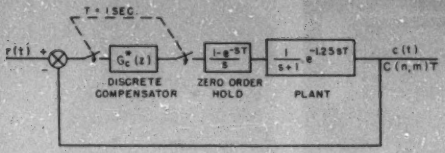


Fig. 10. System used in illustrative example

termined as follows.

The complete response is given by[1,2]

$$C(n,m)T=\frac{1}{2\pi j}\int_{\Gamma}R^*(z)\frac{G_c^*(z)}{1+G_c^*(s)G^*(z)}\times G^*(z,m)z^{n-1}dz \quad (52)$$

For a step input and the given plant, equation 52 becomes

$$C(n,m)T=\frac{1}{2\pi j}\int_{\Gamma}\frac{z}{z-1}\times\frac{G_c^*(z)}{1+G_c^*(z)z^{-1}\dfrac{0.528z+0.104}{z-0.368}}\times\frac{z-1}{z^2}\left[\frac{1}{z-1}-\frac{z^{-m}}{z-0.368}\right]z^{n-1}dz \quad (53)$$

which yields the response at all time, and the response at the sampling instants is obtained for m equal 0.75. If $G_c^*(z)$ is chosen so that the error to a step input is zero for all time after a finite settling time and this settling time is made as small as possible with a physically realizable compensator, then

$$G_c^*(z)=\frac{(z-0.368)z^2}{(z-1)(z+0.21)(z+0.79)\times 0.632} \quad (54)$$

Substituting equation 54 into 53 and evaluating the integral, the complete response is shown in curve c of Fig. 11. This response is a considerable improvement. Furthermore, by comparing equations 51 and 54, it is seen that the compensator is of the same complexity. Also, if Bergen and Ragazzini's method is used and the delay is increased, the overshoot will increase until the system has sustained oscillations with overshoots of 100 per cent for a delay of 1.62 seconds. Such responses are evidently inferior to the uncompensated response, yet the error at the sampling instants is zero. Thus the z-transform can be very misleading.

Such undesirable overshoots between sampling instants can also exist, even though there is no delay, if the plant transfer function has complex poles.[2] The method of this paper and of reference 3 of the paper will fail to indicate or eliminate these undesirable overshoots.

The same argument also applies for design procedures of the author based on minimizing the error squared at the sampling instants. In the opinion of the discussers, if the squared error is to be employed as a criterion of performance, then the square of the input minus the continuous output (or just the continuous output squared for step inputs) should be used. Again the modified z-transform can serve as a useful tool.

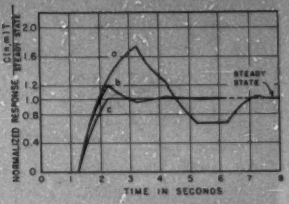


Fig. 11. Actual responses of the plant in Fig. 10

Curve a—Uncompensated response, steady state of 0.5

Curve b—Compensated response using method of the paper

Curve c—Compensated response with use of discussers' method

In conclusion, this paper is further clarification and extension of the work of Bergen and Ragazzini but does not remove the restriction that the response is specified at the sampling instants only, with no modification for critical study between the sampling instants.

References

1. Discrete Compensation of Sampled-Data and Continuous Control Systems. E. I. Jury, W. Schroeder. *Report No. 154*, Electronics Research Laboratory, University of California, Berkeley, Calif., Series 60, 1955.

2. Synthesis and Critical Study of Sampled-Data Control Systems. E. I. Jury. *Ibid., Report No. 136.*

3. Hidden Oscillations in Sampled-Data Control Systems. *Ibid., Report No. 155*, 1956.

An Air-Cleaning Blower for Ventilating Traction Equipment

J. J. GALLAGHER
NONMEMBER AIEE

Paper 56-105, recommended by the AIEE Land Transportation Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Winter General Meeting, New York, N.Y., January 30-February 3, 1956. Manuscript submitted October 20, 1955; made available for printing January 13, 1956.

J. J. Gallagher is with General Electric Company, Erie, Pa.

Problem

Desirability of Clean Air

Reliability of traction equipment on electric and gas-turbine-electric locomotives will be increased, and less maintenance will be required if dirt, rain, and snow can be excluded from the ventilating air. In railroad service brakeshoe dust and water are the greatest enemies of electric equipment; but sand, cinders, and coal dust also cause their share of damage.

An air-cleaning blower has been developed which will economically exclude most of the foreign material from the ventilating air. This reduces the major cause of electrical troubles and insulation failures: the conducting dirts which accumulate in the electric equipment. These dirts build up on the brush-holder studs and the various creepage paths in a traction motor, and when wet, afford very low resistance paths to ground. This

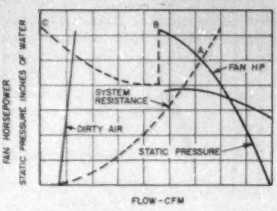


Fig. 1. Characteristic curves of axial flow air-cleaning blower

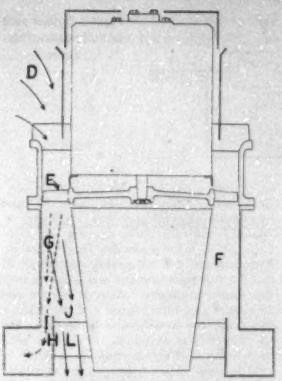


Fig. 2. Cross section of air-cleaning blower showing air-flow paths

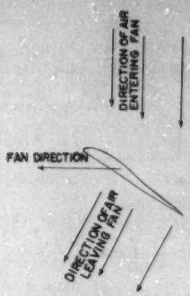


Fig. 3. Diagram of action of fan blade on air stream

tends to produce arcing, with resultant carbonization of the insulation beneath the layer of dirt. In a short time the motor develops a ground and fails in service. Regular maintenance and cleaning programs help to keep the insulation clean and in good electrical condition. Much of the dirt, however, lodges in pockets where it cannot be removed by ordinary cleaning. Obviously, the best method is to remove the dirt and water from the air before it enters the motor.

Since an air-cleaning blower requires only 10 to 15 per cent more horsepower than an equivalent blower without the cleaning feature, its advantages outweigh this slight extra power consumption. When an air-cleaning blower is used the ventilated equipment remains cleaner and drier. Fewer creepage grounds occur and, in general, the insulation remains in better condition. Since the equipment is cleaner there is better heat transfer and the air passages are less likely to become blocked. There is also less damage to the insulation as a result of abrasion.

Basic Requirements

To be effective, an air-cleaning blower must satisfy the following basic requirements:

1. It must have good cleaning efficiencies with all the types of air-borne debris found on a railroad.

2. It must be compact, easy to maintain, and dependable.

3. It must be designed for low cost, high efficiency, and the ability to produce the required air flow and pressure.

There are, in general, two types of fans: centrifugal and axial flow. Each depends on a different principle for operation, and each offers various advantages. A centrifugal fan consists of an impeller which rotates in a scroll-shaped housing. The air enters the revolving impeller at the center or "eye," makes a turn, and passes between the impeller blades. The revolving impeller spins the air between the blades, imparting to it a centrifugal force which expels the air at high velocity, builds up its static pressure, and sucks more air in at the center. Part of the velocity head is converted to static head by the divergence of the scroll-shaped housing from the impeller.

The axial flow fan does not use centrifugal force to produce flow and pressure, but depends upon the screw action of the blades to speed the air along. The air follows a helical path as it is discharged, unless guide vanes are used. These guide vanes turn the air so that the resultant direction of flow of the air leaving the blower is more nearly parallel to the axis of the blower shaft. The guide vanes permit the fan to develop higher pressures by reducing the rotational losses. Considered here are the so-called vane axial fans which have stubby airfoil blades mounted on a hub and rotating in a cylinder. The hub is directly connected to the motor which, in turn, is supported in the cylinder by arms.

Type of Equipment

Advantages and Limitations

An axial flow fan was chosen because it best fits the basic requirements for an air-cleaning blower as outlined. Investigations made during the design of a large electric locomotive showed that compared to a centrifugal fan of equal performance, the axial flow fan would occupy one sixth the cubic space, have two fifths the weight, and cost one half as much. Because of its shape, the axial flow fan lends itself to the use of a one-piece aluminum casting, rather than the welded or riveted construction necessary in a centrifugal blower.

Along with these advantages the axial flow fan also possesses some limitations. Because of its pressure–volume characteristics, such a fan must be matched very closely to the system resistance. The characteristic curve of a typical application is shown in Fig. 1. The operating point A must be near the peak pressure B to obtain maximum efficiency, and yet far enough from the unstable range of the curve from B to C to prevent "stalling" the fan. Stalling is a condition where the quantity of air flow causes the lift coefficient of the airfoil blades to rise beyond their maximum value. This causes the pressure alternately to rise and fall, with a corresponding increase and decrease in the rate of air flow.

The high speed which helps give the axial flow fan its high output per unit of space also causes a high noise level. A check on the noise spectrum of a blower with 24 blades operating at 4,800 rpm indicated that the major portion of the noise occurs at blade frequency and twice blade frequency, in this case approximately 1,920 and 3,840 cycles per second respectively. This is near the middle of the audio range. The noise level may be reduced to some extent by the use of silencers, but it is still higher than that of a comparable centrifugal fan.

Description

The air-cleaning blower is a modification of the standard axial flow traction motor blower used on many General Electric locomotives and on Alco road-type diesel-electric locomotives. The air-cleaning feature depends upon inertia to separate the foreign particles from the air. The blower is mounted in the equipment

Fig. 4. Typical air-cleaning blower for traction-motor ventilation

Fig. 5. Installation of air-cleaning traction-motor blower on gas-turbine-electric locomotive

Fig. 6. Air-cleaning-type blower installed on rapid transit car for ventilating accelerating resistors

compartment as close to the traction motors as possible. In operation, the blower draws air through louvers in the side of the locomotive cab. This air contains all the air-borne debris found in the vicinity of a railroad. Depending on weather conditions, rain or snow may be drawn through the louvers because of the cab vacuum. Drops of water and snow particles will then flow toward the inlet of the blower.

Air containing various quantities of dirt, water, and snow enters the inlet shroud of the blower at *D*, see Fig. 2, and approaches the fan parallel to the axis of the machine. Passing through the fan *E*, the pressure of the air is increased and its direction of flow changed. See Fig. 3. This change in direction causes the dirt-laden air to take a spiral path through the air cleaner *F*. The spinning motion of the air causes the particles of dirt and water, which are heavier than air, to be thrown to the outside by centrifugal force. The path of such particles is shown by the dotted lines at *G*. This has the effect of concentrating the dirt and water at the outer diameter of the spinning air stream while the air towards the center is relatively clean. At *H* a cylindrical baffle divides the air stream. From 12 to 16 per cent of the air containing the highest concentration of dirt flows through the slot at *I* and is collected in a shroud and discharged outside the locomotive cab. The remaining air flows through the opening at *J*, then through the turning vanes *L*, and is discharged into the ventilating duct. The turning vanes change the air direction from a spiral path to an axial path, thus greatly reducing the losses and improving the performance of the blower. The clean air is directed through the duct to the traction motor, control compartment, or other device to be ventilated.

Applications

A typical example of the axial-flow air-cleaning type of blower is shown in Fig. 4. It is the traction motor blower used on the latest gas-turbine-electric locomotives built for the Union Pacific Railroad. The blower will deliver 6,200 cubic feet per minute (cfm) at a static pressure of 10 inches of water while operating at 4,800 rpm and consuming 19 horsepower. Approximately 1,200 cfm of air scavenges the dirt and water to be discharged outside the locomotive. The foregoing table gives the air-cleaning efficiencies attained with various material.

Material	Cleaning Efficiency, Per Cent
Coarse sawdust	92
Fine sand	75
Cinders	72
Water spray	99

To date, air-cleaning blowers of this type have been applied on gas-turbine-electric locomotives for the Union Pacific Railroad, Fig. 5; on the rectifier-type electric locomotives for the New York, New Haven and Hartford Railroad; and on the rapid transit subway cars for the New York Transit Authority, see Fig. 6.

Axial flow air-cleaning blowers of another type were previously applied to several electric locomotives, including the 6,800-horsepower motor-generator locomotives for the Virginian Railway, and the 5,000-horsepower a-c freight locomotives for the Pennsylvania Railroad. Even though these earlier design blowers had a cleaning efficiency approximately 15 per cent less than that of the present design, they have shown excellent results.

The air-cleaning blowers on the Virginian Railway locomotives have now been in service for approximately 8 years. Examination of the traction motors and motor-generator sets shows that they have remained relatively clean. This has resulted in better reliability and reduced maintenance cost. While air cleaning is not a complete remedy, it is definitely an advance in the art of designing and building ventilating equipment for electric apparatus in traction service.

Influence of Atmospheric Water Vapor on High-Current D-C Arcs

R. H. BENNER, II
NONMEMBER AIEE

T. B. JONES
MEMBER AIEE

Synopsis: To determine the effect of water vapor on arc characteristics, arcs with direct currents in the range from 20 to 200 amperes were studied at atmospheric pressure in air containing water vapor in controlled quantities. The water vapor content of the air was varied from 0.15 to 100% (per cent) by volume and this range of values constituted conditions from extremely dry air to saturated steam. For arc electrodes, the materials employed were iron, copper, tungsten, and graphite. Data were taken relating the electrical characteristics and stability of the arc to atmospheric water vapor content, arc current, and electrode separation and material. It was found that the presence of water vapor causes an increase in arc voltage compared to the arc voltage for dry air. This increase is much greater for the electrode materials, iron and copper, than for tungsten or graphite. For a 120-ampere arc with iron electrodes at 1-centimeter separation in an atmosphere of 100% water vapor, the magnitude of the arc voltage is approximately 170% greater than for such an arc in air containing 0.15% water vapor by volume. Whereas, under similar conditions except with graphite electrodes, only a 13% voltage increase occurs. These voltage increases appear to occur in both the arc column and the electrode drop regions. From an analysis of the results, it appears that the differences between air and water vapor effects on the arc derive from the predominant roles of the large specific heats of water vapor and hydrogen and of the great energy transferring properties of hydrogen attributable to its relatively small mass.

THIS paper reports the results of research[1] conducted to determine the effects of atmospheric water vapor on the characteristics of high-current d-c arcs at atmospheric pressure and in air containing water vapor in controlled quantities covering the range from extremely dry air to saturated steam. The objects of this research were to obtain quantitative information to apply to practical problems and to add to the basic knowledge on high-current arcs in gas mixtures.

This work was performed as a related part of an extensive arc study program which has been in progress for some time at The Johns Hopkins University Department of Electrical Engineering. Some of the results of the part of this program dealing with arcs in inert gas mixtures have been presented previously.[2]

Prior work of others[3-8] furnished only a limited guide as to the quantitative water vapor effects on arcs with currents in the primary welding range of 20 to 200 amperes. In brief, these investigators reported that iron electrodes in the current range from approximately 1 to 30 amperes burned with a higher voltage in moist air than in dry air and that an iron electrode arc in a steam atmosphere has electrical and spectroscopic characteristics which are essentially the same as those of an iron electrode arc in a pure hydrogen atmosphere.

As the results of others have shown and those herein will show, water vapor causes an increase in arc voltage. Therefore, the need for an accurate evaluation of this effect is evident, e.g., in the design of electrodes and power supplies for arc welding, automatic electrode feed mechanisms where the control action is determined by the arc voltage, searchlights, projectors, and any other uses of arcs in free atmosphere. In general, past engineering design of equipment and materials involving arcs and related water vapor effects has been done on a highly qualitative and empirical basis. In the practice of arc welding, it has long been recognized that the presence of water vapor can affect the welding process; and, in fact, the specifications for some electrode coatings include empirically established requirements in regard to water vapor generating ingredients. Further, it appears from welding experience in highly humid climate that such climate can constitute poor welding conditions.

The results of the investigation described in the following possibly will contribute to a better understanding of the practical problems involved in this area.

Paper 56-109, recommended by the AIEE Electric Welding Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Winter General Meeting, New York, N. Y., January 30-February 3, 1956. Manuscript submitted October 20, 1955; made available for printing January 13, 1956.

R. H. Benner, II, is an instrumentation and automation consultant, Erlton, N. J., and T. B. Jones is with The Johns Hopkins University, Baltimore, Md.

Experimental Procedure

In the experiments, the controlled arc factors were atmospheric water vapor content, temperature and pressure, electrode material and separation, and arc current. The observed arc factors were arc voltage and stability.

Atmospheric Water Vapor

Atmospheric water vapor content is expressed here in terms of per-cent water vapor per-unit volume. A comparison between these units and those of the concepts of absolute and relative humidities is provided by the several abscissa scales of Fig. 1. For example, the value of 6.4% water vapor by volume corresponds to 100% relative humidity on a basis of 100 F (degrees Fahrenheit) temperature.

The range of water vapor quantities studied was from 0.15 to 100% water vapor by volume. This range was covered by two groups of tests, namely, the low-range and high-range tests. The low-range tests extended from 0.15 to 17% water vapor and covered approximately 1.5 times the maximum probable climatic quantities that occur. The high-range tests extended these studies to 100% water vapor and were conducted to provide a broad base of data for facilitating the analysis of water vapor effects.

The water vapor content of the arc atmosphere was determined in the case of the smallest quantities to a precision of better than 0.1% by volume, and in the case of the large quantities to better than 2.0% by volume.

Atmospheric Pressure and Temperature

Natural ambient atmospheric pressure was used. Control of temperature values of the atmosphere supplied to the arc were studied over the range of from 79 F to 212 F.

Electrode Material and Separation

Electrode materials studied were iron, copper, tungsten, and graphite. Specifications for these were as follows: for iron electrodes: 5/16-inch-diameter bare iron welding rods containing not more than 0.6% carbon; copper electrodes: 1/4-inch-diameter electrolytic copper rods: tungsten electrodes: 1/4-inch-diameter rods of the type supplied for inert gas arc welding; and graphite electrodes: two sizes, 3/8 inch and 1/4 inch in diameter, regular spectroscopic grade rods. Electrode separations from 0.5 to 2.0 centimeters were studied.

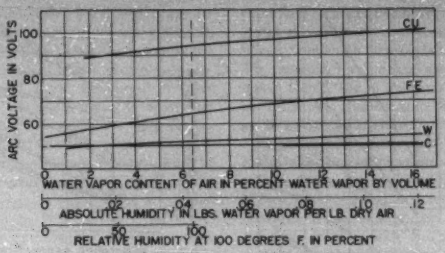

ARC VOLTAGE IN VOLTS
100
80
60
CU
FE
W
C
0 2 4 6 8 10 12 14 16
WATER VAPOR CONTENT OF AIR IN PERCENT WATER VAPOR BY VOLUME
0 .02 .04 .06 .08 .10 .12
ABSOLUTE HUMIDITY IN LBS. WATER VAPOR PER LB. DRY AIR
0 50 100
RELATIVE HUMIDITY AT 100 DEGREES F. IN PERCENT

Fig. 1. Arc voltage versus water vapor content of air arc atmosphere, low range of water vapor values

Arc current—120 amperes
Electrode separation—1 centimeter

Curve Symbol	Material	Diameter, Inches
CU	copper	1/4
FE	iron	5/16
W	tungsten	1/4
C	graphite	1/4

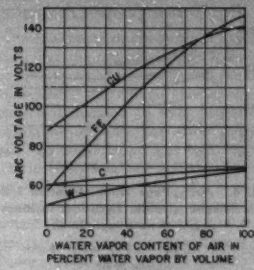


Fig. 3. Arc voltage versus water vapor content of air arc atmosphere, high range of water vapor values

Arc current—120 amperes
Electrode separation—1 centimeter

Curve Symbol	Material	Diameter, Inches
CU	copper	1/4
FE	iron	5/16
W	tungsten	1/4
C	graphite	1/4

Arc Voltage

Voltage across the arc was measured by placing a probe on each electrode at the end away from the arc. Correction for electrode voltage drop was not made since, with low electrode resistances, it was thought that changes in this voltage would be small enough so that their effect on the relative magnitudes of the date would be negligible. The voltage data which are presented represent the average values of numerous test readings.

Arc Stability and Arc Currents

Arc stability was observed by high-speed motion-picture photography made at a rate of 3,000 frames per second. Arc currents from 20 to 200 amperes d-c were studied.

Apparatus

The apparatus was designed to allow careful control of the arc variables treated as test parameters. Fig. 2 shows a general view of the apparatus. This view includes the arc test chamber with back removed, which was 3 feet in diameter and was used for tests in atmospheres containing the low-range quantities of water vapor, the smaller 6-inch-diameter arc test chamber, which was used for atmospheres containing the high-range quantities of water vapor, and part of the equipment for control of the water vapor content and temperature of the atmosphere supplied to the arc.

Arc Atmosphere

Atmospheres containing low - range quantities of water vapor were obtained as follows: Room atmosphere was blown through a duct into a steam-heated, automatically controlled heat exchanger which heated the atmosphere to the temperature controller set point. From there, the heated atmosphere was passed through an automatically controlled steam injection station which injected water vapor into the atmosphere in an amount equal to the water vapor controller set point. From there, the conditioned atmosphere was passed over the sensing elements of the temperature and water vapor control loops. Signals from these elements were transmitted to their respective associated controllers. Next, the atmosphere was passed through venturi tube and hygrometer apparatus for the determination of volume, temperature, and water vapor content. From there, the atmosphere entered the arc chamber at a speed of 115 feet per minute and at a volume of 90 cubic feet per minute.

Atmospheres containing high-range quantities of water vapor were obtained as follows: Compressed air was heated by gas, then mixed with steam, and then passed into the high-range test chamber. Water vapor content was determined from simple gas-law calculations with the use of determined temperature, pressure, and volume values of a sample of conditioned atmosphere, coupled with temperature, pressure, gas volume, condensed water volume, and saturated water vapor volume values determined after cooling the sample to a convenient temperature below the boiling point of the water.

Fig. 2. General view of apparatus

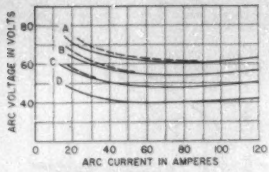


Fig. 4. Arc voltage versus current for 3/8-inch-diameter graphite electrodes in atmospheres of water vapor in air

Solid line curves—2.8% water vapor by volume in air

Dashed line curves—16.2% water vapor by volume in air

Curve Symbol	Separation, Centimeters
A	2.0
B	1.5
C	1.0
D	0.5

Arc Voltage and Current and Power Supplies

Arc voltage and current were measured on strip chart recording meters. Power was obtained through a d-c welding generator or from banks of lead storage batteries or from a combination of both, depending on the arc voltage requirements.

Electrode Feed and Separation

Electrodes were fed by motors with speed controls operated by hand. Electrode separation was determined by viewing the arc image as projected onto a screen marked with a length scale.

Effect of Water Vapor on Total Arc Voltage

Fig. 2 shows arc voltage as a function of atmospheric water vapor in the low range for arcs with 120-ampere current, 1-centimeter electrode separation, and electrode materials, iron, copper, tungsten, and graphite. This shows that arc voltage is increased by an increase in atmospheric water vapor. This effect is greatest for iron electrodes, in which case the arc voltage at 6.4% water vapor is approximately 10 volts, or 18%, greater than at 0.15% water vapor. For graphite electrodes, the effect over this range of water vapor values is essentially negligible.

Fig. 3 shows the same type of curves as Fig. 2 except for the high-range water vapor tests. This shows that arc voltage is increased by an increase in atmospheric water vapor throughout the entire range and that, for an arc with iron electrodes, the voltage for 100% water vapor is approximately 170% greater than for 0.15% water vapor. The effect with graphite electrodes, although causing approximately a 13% total voltage increase for the full range of water vapor values, is still relatively small. Table I summarizes the water vapor effects on total arc voltage for Figs. 2 and 3.

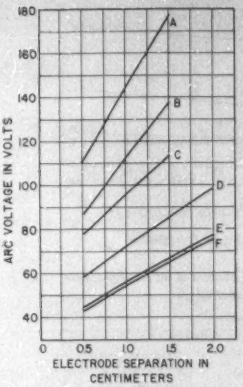


Fig. 5. Arc voltage versus electrode separation for 5/16-inch-diameter iron electrodes in atmospheres of water vapor in air

Arc current—120 amperes

Curve Symbol	Percent Water Vapor by Volume in Air, %
A	100
B	50
C	36
D	16.2
E	2.8
F	0.15

Fig. 4 shows voltage–current curves for arcs with graphite electrodes at separations from 0.5 to 2.0 centimeters and for 2.8 and 16.2% atmospheric water vapor. With such electrodes, an increase in arc current or decrease in electrode separation diminishes water vapor effects on arc voltage. Tests with iron or copper electrodes showed water vapor effects for these low melting point materials to be essentially independent of current.

Table I. Water Vapor Effects

Material	Voltage Increase for Water Vapor Increase from 0.15% to 6.4% by Volume: Volts	%	Voltage Increase for Water Vapor Increase from 0.15% to 100% by Volume: Volts	%
Iron	10	18	93	170
Copper	9	10	54	63
Tungsten	3	6	18	36
Graphite	0.5	1	8	13

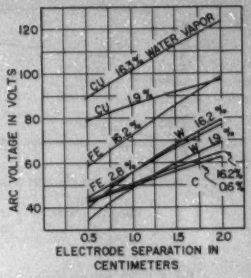


Fig. 6. Comparison of voltage versus electrode separation curves for four electrode materials in atmospheres of air containing water vapor in percentages by volume as indicated on each curve

Arc current—120 amperes

Curve Symbol	Material	Diameter, Inches
CU	copper	1/4
FE	iron	5/16
W	tungsten	1/4
C	graphite	1/4

Within the limits of the accuracy of the observations, tests with atmospheric temperatures varied from 79 to 212 F indicated water vapor effects on arc voltage to be independent of such temperature variations.

Effect of Water Vapor on Arc Column Gradient and Anode-Cathode Drops

Fig. 5 shows voltage–electrode separation (V-S) curves for arcs with iron electrodes, 120-ampere current and atmospheric water vapor values from 0.15 to 100% by volume. Viewing the slopes of these curves as approximating the magnitudes of the arc column gradients, the column gradient is increased by an increase in water vapor. These slopes increase at a rate of approximately 0.5 volt per centimeter for each per cent by volume increase in water vapor. This rate is equal to 2.4% of the slope for 0.15% water vapor. The respective slopes of these curves for 0.15 and 100% water vapors are approximately 21 and 67 volts per centimeter. Straight-line extrapolations of these curves intersect the zero ordinate at values which increase with water vapor. Although the actual shape of these V-S curves at small separations is not known, if the value of the zero intercept is viewed as approximating the magnitude of the sum of the anode and cathode drops, it appears that this sum is increased by an increase in atmospheric

water vapor but to a lesser percentage degree than the column gradient. However, it should be noted that these results are stated in terms of length of electrode separation, and that the actual average length of the arc column is somewhat greater than the electrode separation. Consequently, the slopes and zero ordinate intercepts of these V-S curves will be somewhat greater than the actual column gradients and the sums of the anode and cathode drops respectively. Nevertheless, the percentage changes in the curve slopes and intercept values probably closely represent the respective percentage changes in the column gradients and sums of the anode and cathode drops. Fig. 6 shows a comparison of the same type of results as Fig. 5 for the electrode materials, iron, copper, tungsten, and graphite.

Effect of Water Vapor on Arc Stability

From high-speed motion pictures, an arc with iron electrodes at 1-centimeter separation burns in 100% water vapor with much greater noise and violence of motion and with smaller molten electrode globules than when in 0.6% water vapor. However, it is not difficult to maintain a continuous arc in 100% water vapor if the power supply open-circuit voltage is adequate.

Analysis of Results

Water Vapor Effects on Arc Voltage

Atmospheric water vapor effects on arc voltage can be analyzed in part in terms of arc energy balance. To sustain the arc region in an arc conductive state, energy received from the circuit power supply must equal energy dissipated by the arc processes. Water vapor and its component hydrogen are, in the net effect, less efficient than air as arc gases, primarily because water vapor and hydrogen have relatively large specific heats which are respectively on the order of 2 and 15 times greater than that of air, and because hydrogen has a low mass and, consequently, a relatively high average thermal velocity which is on the order of 4 times that of oxygen and nitrogen at essentially any given temperature. Specific heat is a factor in determining the energy required to increase the temperature of the atmospheric gases convecting and diffusing into the arc to arc temperature; and the small mass of hydrogen contributes to a relatively large rate of energy transfer from the arc region in all directions. Therefore, the energy required to sustain an arc in an atmosphere containing water vapor will be greater than that for an arc in dry air. This increase in energy is obtained through an increase in arc voltage. For example, for the four electrode materials tested, Fig. 6 indicates this result for the column gradient and the sum of the anode and cathode drops.

In addition, however, part of the increase in the sum of the anode and cathode voltage drops may be caused by increases in the effective ionization potentials of the mixtures of electrode vapors and other arc gases in the anode and cathode regions, and such increases may be attributable to greater penetration by hydrogen into these regions. Voltage and spectroscopic observations[6,7] confirm that hydrogen is active in an iron electrode arc in water vapor by showing that such an arc burns essentially as though it were in pure hydrogen.

Influence of Electrode Vapor

Compositions of the mixture of active arc gases is influenced by the degree of electrode vaporization. Iron electrodes and graphite electrodes generate relatively small and large quantities of vapors respectively. Therefore, the ambient arc atmosphere enters substantially into action in the iron arc, whereas, with graphite electrodes, the ambient atmosphere is greatly excluded. The exclusion effect, which increases with increased arc current and diminished electrode separations, is clearly indicated for graphite electrodes by Fig. 4. For bare iron electrodes, however, the effect appears negligible since water vapor effects were found to be essentially independent of arc current.

Influence of Ambient Atmospheric Temperature

Differences in ambient atmospheric temperature in the range from 79 to 212 F have relatively little influence on arc voltage since the differences between gas energies over this range are small compared to the total energies required to raise the temperature of the ambient gases to arc temperature.

Water Vapor Effects on Arc Stability

Violence of the arc activity increasing with water vapor can be ascribed to increased arc voltage or energy dissipation and to the relatively low inertia of hydrogen and, consequently, increased motion of the arc in reaction with the arc convection, electrostatic, and electromagnetic forces. The net effect of the increased arc violence is to reduce the size of the molten metal globules that can form before parting from the electrodes.

Practical Considerations

From a practical standpoint, it appears that, over the range of water vapor contents likely to be encountered in the free atmosphere, the voltage and power increases attributable to water vapor are significant only for the lower melting point metals such as iron and copper. The increases observed with these metals are sufficient to cause difficulty in maintaining an arc for welding purposes in humid climates unless account is taken of this factor by increasing the voltage of the power supply. This need for higher voltage is enhanced by the greater motional activity of arcs in moist air which tends to produce sudden increases in arc voltage requirements, thus leading to instability or loss of the arc unless the power source can supply the necessary voltage.

In the realm of automatic welding, the electrode feed mechanism is frequently of the type which is actuated by the arc voltage. It seems quite possible that changes in the voltage settings on such mechanisms would be necessitated for high ambient humidities to obtain the desired rate of electrode feed. For welding purposes, the increased power dissipated by the arc in moist air is partially recovered by an increased melting rate of the electrodes.

Conclusions

It is felt that the following conclusions may be derived from this investigation:

1. Water vapor in air causes an increase in arc voltage compared to that for dry air.

2. Arc voltage is increased by water vapor much more for arcs with the lower melting point electrodes such as iron or copper than for arcs with tungsten or graphite electrodes. For an arc with 120-ampere current and iron electrodes at 1-centimeter separation in 6.4% water vapor by volume (100% relative humidity at 100 F), the arc voltage is increased to a value of 18% greater than the voltage for extremely dry air (0.15% water vapor by volume). For tungsten under the same conditions, the voltage increase is 6%. In 100% water vapor (saturated steam), the increase in voltage is 170% for iron and 36% for tungsten. Since the currents are equal for these arcs, the increases in arc power required with increased water vapor are directly proportional to the voltage increases. Consequently, power supply equipment for operating arcs in air containing water vapor must have voltage and power ratings increased in proportion to the increases in voltage and power requirements caused by the water vapor above the values for arc operation in dry air. In addition, the voltage settings for electrode feed controls in automatic welding processes would have to be modified in a similar manner to take care of the increased arc voltage when operating under conditions of high atmospheric water vapor content.

3. For all electrode materials tested, the increase of voltage with per cent water vapor follows a somewhat linear relationship; however, the rate of voltage increase becomes progressively smaller with increased water vapor.

4. Arc column voltage appears to be increased more percentagewise than the anode and cathode drops. For iron electrodes, the column gradient appears to increase at a rate equal to approximately 2.4% of the gradient for 0.15% water vapor for each unit percentage increase in water vapor.

5. Differences between ambient atmospheric temperatures in the range from 79 to 212 F have no observed effect on arc voltage.

6. Energy required to sustain an arc appears to be increased by water vapor primarily because of the relatively large specific heat- or energy-absorbing properties of water vapor and its hydrogen component, and the relatively large average velocities and, thus, inherent large energy-transferring properties of hydrogen.

7. To the degree that electrode vapor is generated, it appears to shield the arc from water vapor effects. Vapor from graphite electrodes appears to have a large degree of shielding action which increases correspondingly with an increase in the volume of electrode vapor produced which, in turn, results from an increase in arc current. On the other hand, bare iron electrodes at 1-centimeter separation do not have an observed increase in shielding action as a function of arc current; thus the effect of water vapor in this case is essentially independent of arc current.

8. Arc motional violence and sound are increased, molten electrode globule size is reduced, and the arc column becomes narrower and brighter as atmospheric water vapor is increased. However, if sufficient supply voltage is provided, the ease of sustaining a continuous arc appears to be unaffected by water vapor in the atmosphere.

While this work has thrown some light on the effects of atmospheric constituents on arc behavior, much remains to be learned concerning the detailed mechanisms by which gases and gas mixtures influence the characteristics of high-current arcs. Additional research on arcs with mixtures of the inert gases which is now in progress at Johns Hopkins University holds promise of leading to a better understanding of this problem.

References

1. Influence of Water Vapor on High Current Arcs in Air, R. H. Benner, II. *Ph. D. Dissertation*, The Johns Hopkins University, Baltimore, Md., 1951.

2. Electric Arc and Resistance Welding—III. *AIEE Special Publication S-46*, "Characteristics of the Tungsten Arc in Argon-Helium Mixtures," M. Skolnik, T. B. Jones. Oct. 1952, pp. 198–208.

3. Researches in Arc Welding, G. E. Doan, J. L. Myer. *Electrical Engineering*, vol. 51, Sept. 1932, pp. 624–27.

4. Some Experiments with Arcs Between Metal Electrodes, O. M. Schrum, H. G. Wiest, Jr. *AIEE Transactions*, vol. 50, June 1931, pp. 650–56.

5. Experiments with Short Arcs, G. M. Schrum, H. G. Wiest, Jr. *Electrical Engineering*, vol. 50, Oct. 1931, pp. 827–29.

6. Arcs in Water Vapor, A. Lotz. *Zeitschrift für Technische Physik*, Leipzig, Germany, vol. 15, 1934, pp. 187–91.

7. Voltage Drop of an Arc in Water, A. Lotz. *Wissenschaftliche Veröffentlichungen aus den Siemens-Werken*, Berlin-Siemenstadt, Germany, vol. 14, no. 1, 1935, pp. 25–29.

8. Electrode Consumption Due to Arc Discharge, S. Hoh. *Electrotechnical Journal*, Tokyo, Japan, vol. 3, no. 5, 1939, pp. 99–103.

Magnetic Logic Circuits for Industrial Control Systems

W. G. EVANS, ASSOCIATE MEMBER AIEE — W. G. HALL, ASSOCIATE MEMBER AIEE — R. I. VAN NICE, ASSOCIATE MEMBER AIEE

THE ELEMENTS of automatic control systems may generally be put into three functional categories: 1. Sensing or input elements, which convert external events into useful control signals. 2. Computing elements, which determine the mode of operation of the controlled system on the basis of input information. 3. Machine actuating elements, which convert the electric signals received from the computing elements into actual physical operations. Control systems can be further classified into two general types, analogue and switching. In analogue systems, input signals and output quantities are continuously variable. In switching systems, information is represented by discrete states of electrical or mechanical quantities. Most switching control systems employ a binary representation, in which only two signal levels are distinguished; this permits the use of components which have only two stable states. Relay control systems are the most common binary switching control systems now in use, and information is represented by the two distinct states of the relay contacts. The purpose of this paper is to describe new nonmechanical magnetic elements for use in binary switching control systems.

Experienced relay control system designers think of control problems in terms of relay coils and contacts. These problems can also be thought of in terms of Boolean algebra, as shown by Shannon.[1] In Boolean algebra the action of relays and other binary switching elements are described in terms of the logical functions AND, OR, and NOT. An element which performs an AND function is defined as follows: Signals must exist on all input terminals simultaneously in order for an output signal to be produced. Such an AND element is equivalent to relay "make" contacts in series. An OR element yields an output if signals exist on one or more or its input terminals, and this element is equivalent to relay make contacts in parallel. A NOT element yields an output if no signal exists on its input terminals, and no output is available if an input signal does exist. This element is equivalent to a relay "break" contact. Binary switching control elements also perform the function of memory or storage which is equivalent to a latching relay, and the function of delay which is equivalent to a time delay relay.

As relay control systems become more complex and employ larger numbers of components, the rate of system failures increases. It has long been recognized that a reliable means of performing the functions of relays is needed. The magnetic logic circuits described in this paper provide the desired reliability, and the fast speed of response of such devices may make possible the construction of control systems with capabilities beyond those of present relay control systems.

Paper 56-91, recommended by the AIEE Industrial Control Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Winter General Meeting, New York, N. Y., January 30–February 3, 1956. Manuscript submitted October 21, 1955; made available for printing January 13, 1956.

W. G. Evans and R. I. Van Nice are with Westinghouse Electric Corporation, Pittsburgh, Pa., and W. G. Hall, formerly with Westinghouse Electric Corporation, Jersey City, N. J., is with AMI Inc., Grand Rapids, Mich.

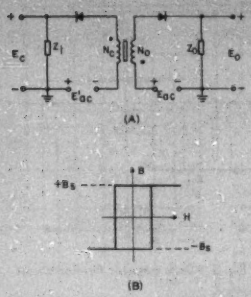


Fig. 1. Basic magnetic amplifier circuit

Background

Circuits that perform the functions of amplification, AND, OR, and memory have been described by Ramey.[1] These circuits are extensively reviewed here to present a basis for the recently developed circuits discussed in a later section.

Amplifier

The basic Ramey magnetic amplifier, which is characterized by a fast time response following a step change in input, is shown in Fig. 1(A). The amplifier consists of two electric circuits linking a saturable magnetic core. The output circuit contains an a-c voltage source E_{ac}, a rectifier, and a path for magnetizing current Z_0, all in series with the output winding N_0. The control circuit contains an alternating voltage source E_{ac}', a rectifier, and a path for exciting current Z_1 all in series with the control winding N_c. E_c is the control voltage. E_{ac} and E_{ac}' have amplitudes proportional to the turns ratio of the two windings, and are in phase with relative polarities as shown in Fig. 1(A). The control voltage E_c, is assumed to be a full-wave rectified sinusoidal voltage in phase with E_{ac} and with the indicated polarity. The dots, Fig. 1(A), refer to winding polarity.

The saturable magnetic core employed in the circuit is of the rectangular hysteresis loop variety having a B-H characteristic approaching that shown in Fig. 1(B). A winding on such a core exhibits a very high impedance when the core flux density lies between the saturation values $\pm B_s$ and a very low impedance when it is attempted to increase the magnitude of core flux density beyond B_s.

In describing the operation of this circuit, it is convenient to distinguish between the two half-cycles of the alternating voltage source. The half-cycle during which E_{ac} has the polarity shown in Fig. 1(A) is termed the output or "gating" half-cycle, and the half-cycle during which E_{ac} has the opposite polarity is termed the control or "resetting" half-cycle.

During the gating half-cycle, voltage is applied to the output winding in such a direction as to drive the core flux density toward positive saturation. The magnitude of E_{ac} is adjusted so that, if the core is at negative saturation at the beginning of the gating half-cycle, it just reaches positive saturation at the end of the half-cycle. During the gating half-cycle the control circuit rectifier has blocking voltage applied to it from the voltage E_{ac}', and possibly the control voltage E_c. These voltages block current flow due to any voltage induced in the control winding from the output winding.

During the resetting half-cycle, E_{ac} and E_{ac}' reverse polarity from that shown in Fig. 1(A). E_{ac}' is then of such polarity as to drive the core flux density toward negative saturation. E_{ac} in the output winding blocks current flow due to voltage induced in the output winding from the control winding.

During the resetting half-cycle, if E_c is zero, E_{ac}' drives the core toward negative saturation, and during this half-cycle the output voltage E_0 is zero. Exciting current flows through winding N_c, the rectifier, impedance Z_1, back to E_{ac}'. During the next gating half-cycle E_{ac} drives the core toward positive saturation, and E_0 is nearly zero (E_0 is a voltage equal to the exciting current times the impedance Z_0).

If the peak value of E_c is of sufficient magnitude during the next resetting half-cycle to block E_{ac}', the core will not be driven toward negative saturation. E_0 will still be zero during this resetting half-cycle. During the next gating half-cycle the core will not support voltage since it is already at positive saturation, and the full value of E_{ac} appears across the load; E_0 equals E_{ac}.

For steady-state operation two conditions of E_c have been considered (neglecting losses in the rectifiers and the core winding)

$$E_c = 0 \qquad E_0 = 0 \tag{1}$$

$$E_c = E_{ac}' \qquad E_0 = E_{ac} \tag{2}$$

The magnetic amplifier has been described here as an on-off device. Within limits the output of the amplifier can be a linear function of E_c[2].

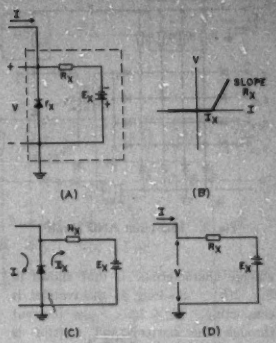


Fig. 2. Nonlinear impedance

When the Ramey amplifiers are used as on-off devices in switching systems, the operation of the systems are not affected by gross changes in the supply voltage, the ambient temperature, or component characteristics. The amplifier is a half-cycle response device, and it can be designed to operate at supply frequencies in the kilocycle range if fast time response systems are desired.

Considerable power gain is afforded by the circuit since only the relatively small magnetizing current of the core flows in the control voltage source, whereas the output current may be many times greater.

The primary function of the elements Z_1 and Z_0 in both the reset and gating circuits of Fig. 1(A) is to provide paths for the flow of magnetizing currents. Ideally, the voltage drop for these currents should be zero, so that the output for zero input will be truly zero, and the full value of the supply voltages used in changing the core flux. An undesirable effect of Z_0 and Z_1 is that they act as loads on the source of input voltage and on the output of the amplifier. Ideally, when the input or output voltage is not zero, no current should flow in Z_1 or Z_0.

If a pure linear resistance R were to be used for Z_1 and Z_0, it is clear that a compromise would have to be made between the two conflicting ideal values of zero and infinity. Such a compromise is possible if nonzero minimum output and a certain amount of loading can be tolerated.

Considerably improved performance can be obtained through the use of the nonlinear element shown in Fig. 2(A). For the case of an ideal rectifier, with no forward drop and no leakage, its volt-

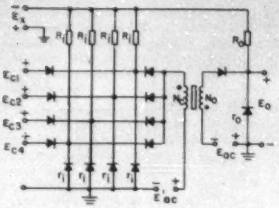


Fig. 3. Four-input AND circuit

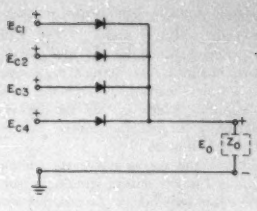


Fig. 4. Four-input OR circuit

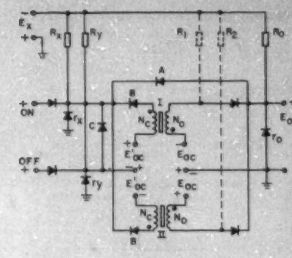


Fig. 5. Basic magnetic flip-flop circuit

ampere characteristic is that shown in Fig. 2(B). So long as the rectifier is conducting, Fig. 2(C), the current through the battery and resistor is $I_x = E_x/R_x$, and the forward current through the rectifier is clearly $I_x - I$, where I is the current from an external source. When I exceeds I_x Fig. 2(D), the rectifier no longer conducts, and I must flow through the battery and resistor, with the resulting drop $IR_x - E_x$ (which is the equation of the line shown in Fig. 2(B)).

The design of a nonlinear element of this type is simple. I_x is chosen to be greater than the magnetizing current to be carried, and R_x is chosen as high as is desirable from considerations of power dissipation, I_x^2R, and loading, $V + E_x/R_x$.

AND-Element

The AND element consists of a basic amplifier of the type shown in Fig. 1(A), with several identical input circuits connected in parallel. Fig. 3 shows a typical four-input AND circuit with the nonlinear elements drawn out. The application of one signal voltage, say E_{c1}, will prevent resetting through that path, but the core will still reset through the others. Only when all four input terminals have signals applied will reset be prevented, and an output obtained. These signals must all occur simultaneously throughout a major part of a reset half-cycle and must have sufficient magnitude to block the resetting voltage E_{ac}'.

OR-Element

The OR element consists of single rectifiers for each unit. Fig. 4 shows a typical four-input OR circuit. In Fig. 4, Z_0 represents the effective input impedance of the element to which the OR circuit is connected. Signals applied to one or more of the input terminals result in a signal voltage across Z_0; i.e., a signal applied to the elements to which the OR circuit is connected. Interaction between the various input signals to the OR circuit is prevented by the rectifiers in series with the inputs.

Memory Element

A memory element, or flip-flop, may be constructed with the use of two of the basic amplifier circuits of Fig. 1(A). The two basic circuits are connected in parallel, as shown by the solid lines of Fig. 5. The supply voltages E_{ac} and E_{ac}' are connected with the phase relationships shown, so that when one core is gating, the other is resetting. The common output voltage is fed back to the common resetting circuit through rectifier A. Resetting of both cores is accomplished through rectifier r_x and rectifier r_y. A direct current flows through r_y in the forward direction and through R_y. Because of the low forward drop of r_y, the center tap of E_{ac}' is essentially grounded.

The instantaneous a-c polarities of Fig. 5 indicate the condition of core I being reset and core II being gated. If core II had been reset during the previous half-cycle, there would be zero output voltage E_0, and hence no voltage fed back through rectifier A. Thus, core I would be reset by E_{ac}'. During the next half-cycle, core I will gate with zero output voltage, and core II will reset. Thus, the output E_0 will remain at zero.

The application of a signal of half-cycle duration to the ON terminal will block rectifier B, and prevent the resetting of whichever core is in its resetting half-cycle. This core will gate during the following half-cycle and output voltage E_0 will appear. This voltage is fed back through rectifier A and will continue to block rectifier B, preventing the resetting of the other core. The circuit thus remains in the full output state, even though the ON signal may be removed.

The introduction of a signal of half-cycle duration to the OFF terminal will block rectifier r_y. This raises the center tap of E_{ac}' above ground by the magnitude of the OFF signal. Thus, the resetting voltage is increased sufficiently to allow rectifier B to conduct and allow one of the cores to reset. During the next half-cycle, that core will gate with zero output voltage, and the other core will be allowed to reset. Rectifier C prevents excessive voltage from possibly being applied from the OFF terminal to the reset windings if there be no ON signal or feedback pulse from rectifier A.

The dotted resistors, R_1 and R_2 of Fig. 5, will be discussed later.

Recent Developments

The circuits reviewed here perform satisfactorily the functions of amplification, AND, and OR. The flip-flop circuit described has disadvantages, and improvements and additions to the basic flip-flop circuitry to overcome its disadvantages and increase its versatility are now described.

Flip-Flops

The previous description of the basic memory or flip-flop circuit used the assumptions that the residual flux of the magnetic core approaches the saturation flux, and that the gating rectifiers do not have leakage in excess of the magnetizing magnetomotive force of the core. If either of these assumptions are not met, the result is a partial reset of the core during those reset half-cycles in which the reset circuit is blocked by either an ON signal, or the feedback voltage. This partial reset leads to an output during the subsequent gating half-cycle. Thus, the feedback voltage does not completely block reset during the next resetting half-cycle, and more reset occurs. This condition can continue to the point where the flip-flop will turn off without the presence of an OFF signal.

The instability just described can be remedied by the addition of resistors R_1 and R_2 of Fig. 5. These resistors allow

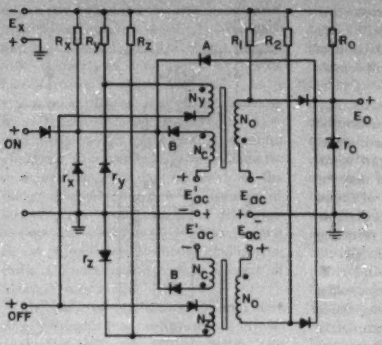


Fig. 6 (left). Three - winding flip-flop circuit

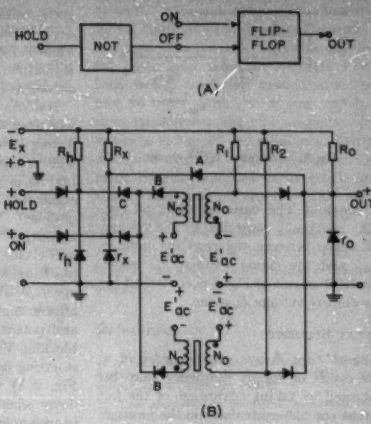


Fig. 7 (right). Flip-flop with hold input

a small amount of direct current to flow through both output windings in such a direction as to hold each core in positive saturation after the cores have been saturated during a gating half-cycle. This direct current is small enough so that it will not drive the cores to positive saturation if the core resets to negative saturation. Thus the output during a gating half-cycle will be a complete half-cycle and will completely block the resetting voltage.

Three-Winding Flip-Flop

The two-winding flip-flop discussed in the preceding paragraphs can be improved upon. Examination of Fig. 5 will show that the resetting supply voltages E_{ac}' must be obtained from a separate supply for each flip-flop. Because the common E_{ac}' terminal is not grounded, if several flip-flops are connected to the same resetting supply, and an OFF signal is applied to only one flip-flop, all the others will also turn off.

One method of overcoming this characteristic is to provide a separate transformer for each flip-flop. Another means is to wind the required number of isolated windings on the main transformer for the system. Obvious disadvantages to both these schemes are the increased winding costs, wiring costs, and greater space requirements. In addition, the type of OFF input circuit shown in Fig. 5 has the disadvantage of drawing more current from the signal source than the ON input does.

These disadvantages to the flip-flop circuit of Fig. 5 may be overcome by the addition of a third winding to each core. This new flip-flop is shown in Fig. 6. In this circuit, the resetting and gating circuits remain unchanged from those of Fig. 5. Thus this circuit turns on as before. However, the OFF signal is now applied to a separate winding, N_y and N_z, on each core. Elements R_y, r_y, and R_z, r_z make up two nonlinear impedances that limit the current through each winding N_y and N_z. The action of an input signal voltage to the OFF terminal is to reset that core which is undergoing its reset half-cycle. The output of this core in the subsequent gating half-cycle is zero, which allows the other core to be reset by E_{ac}'. A continuous OFF signal will reset both cores and prevent any output from the flip-flop no matter if an ON signal is impressed simultaneously.

Hold Flip-Flop

The previously described flip-flops all require the presence of an OFF signal to change their states from ON to OF. There frequently arises a need to turn a flip-flop off by the absence of a signal. Fig. 7(A) shows how this may be accomplished by the use of an auxiliary NOT element. In a NOT element (described later) no output exists when an input signal is applied. If an input signal is provided to the NOT element, the flip-flop can be turned on. As long as the NOT has an input, the flip-flop will remain on. As soon as the NOT input is removed the flip-flop will be turned off. The input to the NOT can be termed a HOLD signal, since its presence holds the flip-flop on. This combination of a NOT and a flip-flop to obtain the desired function is usually expensive because of the added cost of the NOT element.

A more inexpensive and practical means for holding a flip-flop on is shown in Fig. 7(B). The ON and feedback circuits are the same as before. Another resetting path is provided by the addition of R_h, r_h, and C. Thus a signal voltage must be applied to the HOLD terminal to prevent reset of the cores and allow the flip-flop to be turned on and remain on. When the HOLD signal is removed, reset will always occur through the HOLD circuit and the flip-flop will remain off, even though the ON signal is applied.

The flexibility of the flip-flop may be increased by adding both the HOLD input and the OFF input to a third winding. In this case, either the absence of a HOLD signal or the presence of an OFF signal will turn the flip-flop off.

NOT-Element

The circuit of Fig. 8 performs the function of NOT, or negation. The winding polarities (dots of Fig. 8) are such that the signal voltage E_s drives the core to negative saturation, while the supply voltage E_{ac} drives the core to positive saturation. With no input signal E_s, reset of the core is impossible, and full output voltage E_0 will occur during each gating half-cycle. Reset will occur only when a signal voltage E_s of sufficient magnitude to reset the core is applied during the resetting half-cycle of the supply voltage. If this occurs, during the following gating half-cycle no output voltage E_0 will appear. The nonlinear impedance element consisting of E_x, R_i, and r_i limits the resetting current drawn from the signal source.

While this circuit may seem similar to that of Fig. 1(A), it differs in the manner

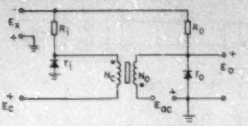


Fig. 8. Magnetic NOT circuit

of control. The basic amplifier of Fig. 1(A) is actually controlled by the voltage difference between the resetting voltage E_{ac}' and the signal voltage E_c, while the NOT circuit of Fig. 8 is controlled by the control voltage E_c alone.

Delay Element

Short time delays on the order of a few cycles of supply frequency may be obtained by taking advantage of the inherent one-half cycle delay in the passage of a signal through the basic amplifier of Fig. 1(A). By cascading a number of amplifier stages, delays of integral multiples of half-cycles may be obtained.

Larger delays up to 10 or 20 cycles can be obtained by modification of the basic amplifier as shown in Fig. 9(A). The control or resetting circuit of Fig. 9(A) is similar to those previously described. The output or gating circuit has several distinct differences from the previously described circuits. The gate winding N_b is designed to support some integral multiple of supply voltage E_{ac}. Fig. 9(B) shows the ideal hysteresis loop for a delay core wound to support ten times the supply voltage E_{ac}. Supply voltage E_{ac}'' of Fig. 9(A) is phased with respect to E_{ac} so that during the gating half-cycle shown, E_{ac}'' forces forward current through rectifier r'', and the components r'' and R'' act as a nonlinear impedance as described previously. Rectifier r'' is biased to pass exciting current and full load current.

If no control voltage E_c is applied to Fig. 9(A), the core will reset to negative saturation during each resetting half-cycle. During each gating half-cycle, voltage E_{ac} will be absorbed by the core, and only the small area number 1, Fig. 9(B), of the hysteresis loop will be traversed. Now if E_c is applied at the beginning of a resetting half-cycle, reset will be blocked, and the core will remain at position 1. During the next gating half-cycle, voltage E_{ac} will again be absorbed by the core. If E_c is still present, the core will remain at 2, Fig. 9(B) during the next resetting half-cycle. The core will continue to absorb E_{ac} each succeeding gating half-cycle until saturation is reached. During the next gating half-cycle, voltage E_{ac} will appear across rectifier r_b and output voltage E_o will exist. Thus 9.5 cycles after the application of E_c, output E_o will appear, and will remain as long as E_c is applied.

If E_c is removed after E_o exists, E_{ac}' will reset the core to negative saturation during the next resetting half-cycle. The voltage induced in the output winding N_o is greater than E_{ac} and may be equal to E_{ac} times the number of cycles of delay. This induced voltage will cause an output to appear during the resetting half-cycle. To prevent an output during resetting, the elements R'', r'', and voltage E_{ac}'' are added. E_{ac}'' is a sinusoidal alternating voltage of such magnitude and polarity as to add E_{ac} in completely blocking the induced voltage during the resetting half-cycle. Thus the circuit of Fig. 9(A) will give a number of cycles delay when E_c is applied, and resetting is accomplished in one half-cycle.

If resetting in one half-cycle is not required, the resistor R'' and rectifier r'' may be eliminated. By proper adjustment of the nonlinear resistance R_l and r_l in the resetting circuit, the induced voltage in the output winding can be limited.

To obtain delays longer than the 10 or 20 cycles obtainable with a single core, several of these elements can be cascaded. By this means delays of seconds or even minutes are possible.

Application of Magnetic Logic Elements

One great advantage of relays is that the input to the coil is isolated from the output or relay contacts. Another advantage is the ease of interconnection of relays. The contacts of a relay can be used to energize many other relay coils. The magnetic logic elements described here offer galvanic isolation between the input and output signals since the input and output windings are separate. Magnetic logic elements are simple to interconnect because a common ground can exist between the input and output windings. The input circuits of all elements can be designed to operate from the output of all other elements. Conversely, the output of any element can be used to drive any other element. The output of one element can be used as inputs to 10 or 20 other elements.

The relay control system designer looks upon a relay as a unit or a building block. The magnetic elements described here are similar self-contained building blocks. Three or more common wires supply power to all units. One common ground

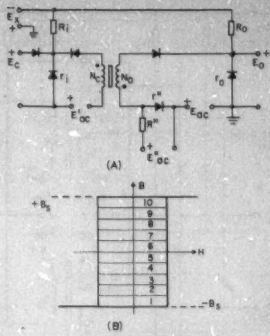


Fig. 9. Magnetic delay circuit

is used for all units. The output of one element is connected directly by a single wire to the input of any other element or several elements. No impedance matching problems exist. No attenuation problems exist if no more than a specified number of elements are driven by any one element. A logic diagram as shown in Fig. 7 becomes a wiring diagram except for the common power leads.

Magnetic amplifiers present problems when subjected to widely varying ambient temperatures in industrial and military applications. Since these amplifiers are usually dependent upon a linear relation between the input and output, changes in the cores or rectifiers cause changes in the performance of magnetic amplifier control systems. Magnetic logic elements are used as on-off devices, and gross changes in components will not affect the performance of systems which employ them. Logic elements containing conventional selenium rectifiers will function satisfactorily in ambients from −60 to +80 degrees centigrade. Normal changes in components with age will not affect the performance of the elements.

Magnetic logic elements may be operated at signal levels in the range common to conventional magnetic amplifier systems: microwatts to kilowatts. The choice of the level will be dependent upon the components available, the amount of complexity permitted, and the economics involved. If input signals are low in power, amplification can be used. Again, amplifier linearity is not a problem because of the on-off nature of the elements. If the output of a given logic element design is not suitable, amplification can be used. If a large amount of computation is re-

quired as compared to the number of high power outputs required, economics and good design practice would dictate logic elements of a milliwatt level with amplification to drive the load. As a typical example, a 1-inch inside diameter by $1^1/_4$-inch outside diameter by 1/4-inch strip width, tape-wound toroidal Hipernik V core, and 1/4-inch-diameter selenium rectifiers can be used in the design of an AND circuit for operation at 60 cycles per second capable of delivering 1/4 watt to a load or other elements for a signal power input of about 8 milliwatts. Such a device would have a volume of about 4 cubic inches.

The ease of application of magnetic logic elements has been stressed. However, the designer will encounter two great differences between relays and magnetic elements. The output of some magnetic elements is half-wave rectified sine waves. This requires that phasing be taken into account when applying the output of one element to another. A relay may have a large number of contacts for each signal coil. A magnetic logic element will have only one output signal for many input signals. Certain techniques of design must be learned in order to replace relay control systems with static systems. A number of industrial applications of magnetic logic elements have been made to date. These elements have been combined to control the movement and welding of bus duct channel in an automatic spot welding machine. All of the usual interlocking and safety features have been incorporated in the design. The system has been in operation in an industrial production line for six months without trouble. Such service represents millions of operations in equivalent relay controls.

Conclusion

Static magnetic devices are available to replace relays in switching control systems. Although some of the circuitry shown here has been developed recently, the basic components, magnetic amplifier cores, and selenium rectifiers have been widely used and proved over the past 10 to 20 years. Tests and applications have shown the new magnetic elements to be rugged and reliable, and they will have long life. The devices can be made on self-contained cards which are simple and easy to apply. The wide tolerances which are permissible in components will allow magnetic elements to compete economically with relays in industrial and military control systems.

References

1. A Symbolic Analysis of Relay and Switching Circuits, Claude E. Shannon. *AIEE Transactions*, vol. 57, 1938, pp. 713–23.
2. The Single-Core Magnetic Amplifier as a Computer Element, R. A. Ramey. *Ibid.*, vol. 71, pt. I, 1952 (Jan. 1953 section), pp. 442–46.
3. On the Control of Magnetic Amplifiers. *Ibid.*, vol. 70, pt. II, 1951, pp. 2124–28.

Negative-Sequence Voltage Detection on Aircraft A-C Electric Systems

R. D. JESSEE
ASSOCIATE MEMBER AIEE

P. F. BOGGESS
ASSOCIATE MEMBER AIEE

SHORT circuits or grounds in aircraft electric power systems or in other parts of the electric system cause unbalanced voltages. These voltages may be detected by several means. One method indicated by a study of symmetrical components is the negative-sequence voltage.[1] It can be shown that line-to-line and line-to-ground fault detection can be provided by sensing the negative-sequence voltage.

Negative-sequence filter circuits have been used for many years in large industrial power systems.[2] An investigation of these circuits related to the aircraft system problems reveals that some form of frequency compensation is required if the circuits are to be applicable to an aircraft electric power system. The aircraft system frequency may normally vary during load transients as much as ±5 per cent and, during system control such as startup and shutdown, by as much as 30 to 50 per cent. The normal negative-sequence voltage-sensing circuit would produce an erroneous signal because of the frequency variation. The authors are presenting a solution to the problem of sensing negative-sequence voltage which includes the provision for frequency compensation.

The circuit will be analyzed showing how its performance may be evaluated and how it may be applied to electric system problems. Included in this evaluation is sufficient circuit analysis to permit study of circuit loading on the sensing circuit output, and consideration of the effect on sensing circuit output of harmonic voltage input.

Paper 56-440, recommended by the AIEE Air Transportation Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE South West District Meeting, Dallas, Tex., April 2–4, 1956. Manuscript submitted January 3, 1956; made available for printing April 23, 1956.

R. D. Jessee and P. F. Boggess are with the Westinghouse Electric Corporation, Lima, Ohio.

Nomenclature

f = per-unit frequency, the ratio of actual to nominal frequency
V_{ab}, V_{bc}, V_{ca} = line-to-line voltages between phases indicated by subscripts
V_1 = positive-sequence voltage of phase a
V_2 = negative-sequence voltage of phase a
V_{C1} = voltage across capacitance C_1
V_{R1} = voltage across resistance R_1
V_{sen} = output voltage of sequence filter network
V_{comp} = output voltage of frequency-compensating circuit
V_{sig} = output voltage of complete compensated circuit (circuit unloaded)
V_T = reference voltage, output of T_1
V_f = output voltage of frequency-sensitive network
π = constant of proportionality R_4/R_3
V_L = output voltage under load
I_L = output load current
Z_f = equivalent impedance of frequency-sensitive network
Z_s = equivalent impedance of sequence filter
Z_L = impedance of load
X_1, X_2 = reactances of C_1 and C_2 respectively
Z_T = impedance of the circuit supplying V_T

The Frequency Problem in Sensing Negative-Sequence Voltage

A simple negative-sequence voltage-sensing circuit is shown in Fig. 1. Line-to-line voltages are applied to this sensing circuit. The phasor relationship of the line-to-line voltages and the voltage components are shown in Fig. 2(A) and 2(B) when rated frequency is applied to

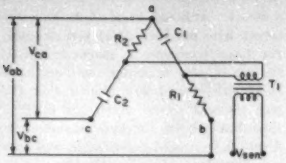


Fig. 1. Negative-sequence filter circuit

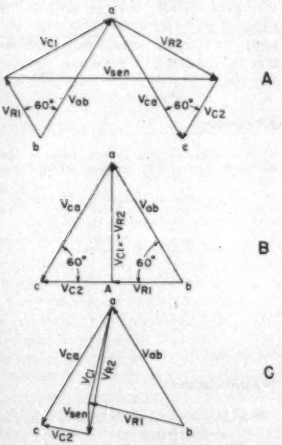


Fig. 2. Phasor diagrams of sensing circuit with V_1 and V_2 input

A—Negative-sequence voltage V_2 input at 1.0 per-unit frequency

B—Positive-sequence voltage V_1 input at 1.0 per-unit frequency

C—Positive-sequence voltage V_1 input at 1.3 per-unit frequency

the network. Fig. 2(A) shows the output of the filter (negative-sequence voltage only) when applied to the sensing circuit. Fig. 2(B) shows that, with only positive-sequence voltage input, the output signal is zero as indicated at point A.

The detailed analysis by phasor relationships is given in the Appendix. The sensing voltage is

$$V_{sen} = V_{R2} + V_{C1} \tag{1}$$

where

$$V_{R2} = \frac{R_2}{R_2 - jX_2} V_{ca} \tag{2}$$

and

$$V_{C1} = \frac{-jX_1}{R_1 - jX_1} V_{ab} \tag{3}$$

Substitution of equations 2 and 3 into

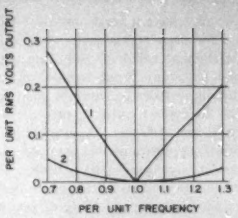


Fig. 3. Output error caused by frequency variation with no negative-sequence voltage input V_2. Curve 1 is the output from the uncompensated circuit, Fig. 1, and curve 2 is the output from the compensated circuit, Fig. 7

equation 1 results in the equation for the sensing voltage

$$V_{sen} = \frac{3}{2} \frac{[f^2-1+j\sqrt{3}(f^2-2f+1)]\,V_1 + [1-f^2+j\sqrt{3}(f^2+2f+1)]\,V_2}{4f+j\sqrt{3}(f^2-1)} \tag{4}$$

The equation for determining the sensing voltage comprises three variables: frequency, positive-sequence voltage, and negative-sequence voltage. By inspection of equation 4, it is noted that, with no negative-sequence voltage V_2 and at nominal frequency ($f=1$), no sensing voltage would appear at the output terminals of the network. The phasor diagram, Fig. 2(B), shows this sensing voltage as being zero (indicated at point A). If the frequency is varied from nominal with only positive-sequence voltage input, a voltage error will be detected at the output. This voltage is shown in Fig. 2(C) as V_{sen}.

The output voltage of the negative-sequence sensing circuit without frequency compensation is shown as a function of frequency in Fig. 3, curve 1. If the operating level were within the magnitudes expressed on this curve, tripping of a relay using this sensing circuit could result with only a change in frequency. Therefore, it is apparent that, if it is desired to measure accurately, the negative-sequence voltage, frequency compensation in some form must be provided.

The Analysis of a Frequency-Compensated Sensing Circuit

The output voltage error caused by a frequency deviation is shown in Fig. 2(C). The error voltage is approximately 90 degrees out of phase with the line-to-line voltage V_{bc}. A curve of the component of error voltage, 90 degrees out of phase with V_{bc}, is plotted as a function

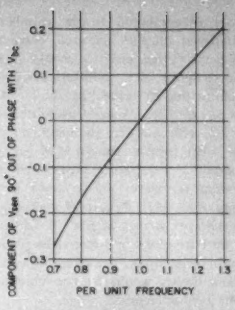


Fig. 4. Error voltage with variation in frequency when positive-sequence voltage is 1.0 per unit and negative sequence voltage is 0.0

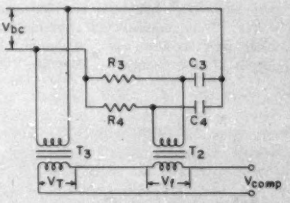


Fig. 5. Frequency-compensating network

of frequency in Fig. 4. Having developed the problem to this point, it is apparent that, if the desired voltage is to be sensed, it will be necessary to provide a frequency-compensating circuit which would nullify the error shown in Fig. 4.

Such a compensating circuit is shown in Fig. 5, and the performance of this circuit is described by the following:

$$V_{comp} = j\sqrt{3}(V_1 - V_2) \times \left(\frac{n^2-1}{n^2+1} + \frac{j(n^2-1)f}{n(f^2-1)-j(n^2+1)f}\right) \tag{5}$$

The output of the circuit described in the foregoing equation is nearly equal and opposite to the error shown in Fig. 4.

The phasor diagrams explaining the operation of the frequency-compensating network are given in Fig. 6. Figs. 6(A) and (B) show the voltage relationships at nominal and higher frequencies, necessary to develop the frequency-sensitive voltage V_f. To obtain a voltage nearly equal and opposite to the voltage error V_{sen} in Fig. 2(C), a reference voltage V_T is combined with the frequency-sensitive voltage V_f. The phasor addition of V_T and V_f to obtain the compensating voltage V_{comp} is shown in Fig. 6(C) where

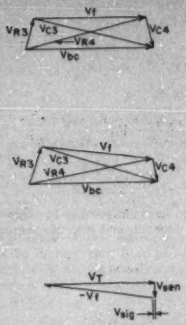


Fig. 6. Phasor diagrams of frequency-compensating network

A—Frequency-sensitive network at 1.0 per-unit frequency
B—Frequency-sensitive network at 1.3 per-unit frequency
C—Phasor addition of frequency-compensating network voltage and sensing-circuit voltage

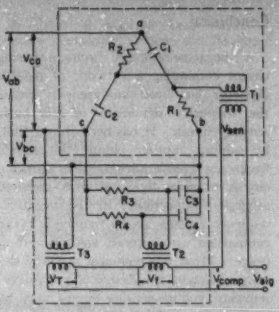


Fig. 7. The circuit diagram of the frequency-compensated negative-sequence voltage-sensing network

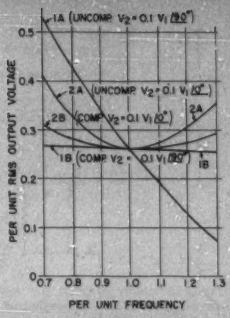


Fig. 8. Output characteristics showing how compensating circuit corrects for frequency variation

$$V_{comp} = V_2 - V_1 \quad (6)$$

The remaining voltage error V_{sig}, shown in Fig. 6(C) is very much reduced magnitude from the original voltage error caused by frequency. The error may be calculated by

$$V_{sig} = V_{sen} + V_{comp} \quad (7)$$

The complete circuit diagram for the negative-sequence voltage-sensing circuit is shown in Fig. 7. Fig. 3, curve 2, describes the error over a frequency range of 0.7 to 1.3 per unit of nominal frequency for the compensated circuit. The reduction of the error by frequency compensation, with no negative-sequence voltage input, is shown by comparison of curve 1 to curve 2 in Fig. 3.

Short circuits or grounds in an electric system will cause unbalanced voltages that produce component voltages, V_1 and V_2, which have an angular relationship one to the other. For example, if V_1 and V_2 are in phase on one phase, V_1 will be displaced from V_2 by 120 degrees on the other two phases. The effect of the phase displacement on the output of the sensing circuit is shown in Fig. 8.

Fig. 8 shows how compensation improves the accuracy of the output voltage when a negative-sequence voltage input is applied. Here curves 1(A) and 2(A) show the output voltage with a constant input voltage $V_1 + V_2$, where V_2 is 90 degrees out of phase with V_1, and where V_2 is in phase with V_1. Curves 1(B) and 2(B) show the output voltages when compensation is applied. It is important that, over a frequency range of ±10 per cent, there is negligible error in this circuit whenever phase displacement is considered, and that appreciable error occurs only at very low or very high frequencies with a frequency-compensated circuit.

Fault Detection by Means of Frequency-Compensated Circuit

The effect of frequency error has been simply shown for a line-to-line fault condition and a line-to-neutral fault condition in Fig. 9. These fault conditions were calculated on the basis of a 40-kva aircraft-type a-c generator operating at 400 cycles per second. The fault values are not shown for frequencies other than 1.0 per unit because the error is small since the fault impedance is large with respect to machine impedance in the operating range. For application as a means of fault detection, a frequency range of ±10 per cent should require no more consideration than the use of the curve where frequency is equal to 1 per unit and the value of fault current is determined at the output level of the sensing circuit. An example calculation is shown in Fig. 9. During startup or shutdown conditions or where application is required over a wider frequency range, other curves may be extrapolated, such as where frequency is shown equal to 0.7 per unit for a line-to-neutral fault or 0.7 per unit for a line-to-line fault. The equations given in the analysis of this device lend themselves to any type of evaluation necessary for system application.

Consideration to Load Effects and Harmonics on Sensing-Circuit Output

The method of analyzing load effects is given in the Appendix. Where high input impedance circuits such as electronic amplifiers are coupled to the

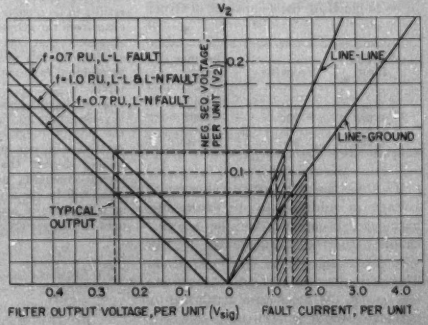


Fig. 9. The range of line-to-line and line-to-ground fault protection provided with a typical negative-sequence voltage filter output. The fault current and negative-sequence voltage data computed on a 40-kva a-c generator regulated for maximum phase voltage of 1.0 per unit

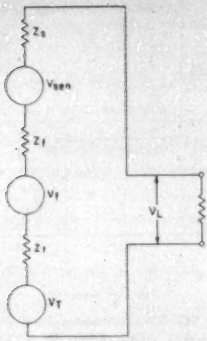


Fig. 10. Equivalent circuit for studying load effects

sensing circuit, no consideration need be given to load effects, but, if this circuit were to be applied to an electromechanical relay, consideration would be necessary to determine circuit constants required for its application.

Harmonic voltages in the electric system must be considered when applying the sensing circuit to a system problem. The nature of harmonic voltages, i.e., whether they are of a positive, negative, or zero-sequence type at some frequency related to the fundamental permits direct use of equation 7 for evaluation. Zero-sequence harmonic voltages need not be considered as they will not pass through the filter. Positive- and negative-sequence harmonics usually need not be considered unless the magnitude of the output caused by these harmonics is relatively high in comparison to the output signal for a prescribed negative-sequence voltage.

The following example in per-unit quantities of fundamental voltages will serve to illustrate the magnitude of error:

Per-Unit Input	Harmonics	Per-Unit Output
0.02	2nd	0.054
0.02	5th	0.047
0.02	7th	0.0055
0.02	8th	0.044

If 0.1 per-unit negative-sequence voltage in addition to the foregoing harmonics is applied to the sensing circuit, the root-mean-square of the voltage output will be 0.273 per unit. Therefore, since 0.26 per unit is the nominal output, the signal error attributed to harmonics would be 5 per cent.

Conclusion

A negative-sequence voltage-sensing circuit employing a network of passive elements has been analyzed, revealing the frequency deficiency of the basic sensing circuit. It has been shown that, by adding another circuit to provide frequency compensation, the resulting circuit provides a signal with required accuracy proportional to the negative-sequence voltage over a frequency range compatible with aircraft electric system usage. From the technical material developed herein, the design of a negative-sequence voltage-sensing circuit may be established, and the resulting design may be evaluated for performance in various aircraft electric system applications.

Appendix. Development of Equations Describing the Operation of the Frequency-Compensated Negative-Sequence Voltage-Sensing Circuit

Sequence Filter

Refer to Fig. 1, the circuit diagram of the sequence voltage filter only. The circuit constants are so chosen that at nominal frequency ($f=1.0$ per unit) the voltage drop across resistor R_3 leads V_{ca} by 30 degrees and the voltage drop across capacitor C_1, lags V_{ab} by 30 degrees. To meet these conditions

$$X_1=\sqrt{3}R_{1/f}$$

$$R_3=\sqrt{3}R_1$$

$$X_3=R_{1/f}$$

It is assumed that the transformation ratio of the isolating transformer T_1 is unity. As shown in Fig. 2(A), a balanced input voltage of negative sequence results in a sensing voltage V_{sen} of $-3/2V_{bc}$. Likewise, it is shown in Fig. 2(B) that a balanced input voltage of positive sequence results in no sensing voltage, $V_{sen}=0$. It is clear from the geometry of these two phasor diagrams that the sequence filter is responsive to only negative-sequence voltages at nominal frequency, fundamental voltage input.

To understand what happens as frequency is varied, refer to Fig. 2(C). Here is shown the phasor diagram of the filter having positive-sequence input only, but at a frequency higher than normal. Using the circuit constants previously stated, equations will be developed for sensing voltage as a function of sequence component voltages and frequency. In Fig. 2(C)

$$V_{R3}=\frac{R_3}{R_3-jX_3}V_{ca} \tag{2}$$

$$V_{C1}=\frac{-jX_1}{R_1-jX_1}V_{ab} \tag{3}$$

$$V_{sen}=V_{R3}+V_{C1} \tag{1}$$

$$=\frac{R_3}{R_3-jX_3}V_{ca}-\frac{jX_1}{R_1-jX_1}V_{ab}$$

In terms of R_1

$$V_{sen}=\frac{R_1\sqrt{3}}{R_1\sqrt{3}-jR_{1/f}}V_{ca}-\frac{1\sqrt{3}R_{1/f}}{R_1-j\sqrt{3}R_{1/f}}V_{ab}$$

$$=\frac{f\sqrt{3}}{f\sqrt{3}-j1}V_{ca}-\frac{j\sqrt{3}}{f-j\sqrt{3}}V_{ab} \tag{8}$$

The line-to-line voltages may be expressed in terms of positive- and negative-sequence components[1] as

$$V_{ab}=(a^2-1)V_1+(a-1)V_2$$

$$V_{bc}=(a-a^2)(V_1-V_2)$$

$$V_{ca}=(1-a)V_1+(1-a^2)V_2$$

By substituting for V_{ca} and V_{ab} in equation 8 and reducing, the equation for V_{sen} may be written as a function of positive- and negative-sequence voltages and frequency as follows

$$V_{sen}=\frac{3}{2}\frac{[f^2-1+j\sqrt{3}(f^2-2f+1)]V_1+[1-f^2+j\sqrt{3}(f^2+2f+1)]V_2}{4f+j\sqrt{3}(f^2-1)} \tag{4}$$

A study of equation 4 will show that, when balanced line voltages exist ($V_2=0$), the sensing voltage V_{sen} stays very nearly in phase with V_1 (or 90 degrees displaced from V_{bc}) as the frequency is varied over the operating range of aircraft generators.

Frequency-Compensating Circuit

The frequency-compensating circuit is shown in Fig. 5. It is so designed as to produce a voltage approximately equal and opposite to the frequency-error voltage produced by the sequence filter. The circuit comprises two voltage sources: one a frequency-sensitive network, the other a reference voltage source. These two voltages are added phasorially and their sum is the compensating voltage. Referring to Fig. 5

$R_4=nR_3$, where n is a constant of proportionality

$X_4=R_{1/f}$, the reactance of C_4

$X_3=nR_{3/f}$, the reactance of C_3

The value of n is somewhat arbitrary. A value of $n=4$ seems reasonable although not necessarily optimum and was used for calculations of curves. It will also be assumed that the transformation ratio of the isolating transformer T_2 is unity.

From these equations, it is evident that at nominal frequency X_3 equals R_4 in magnitude, X_4 equals R_3, and that the voltage V_f, as shown in Fig. 6(A), is in phase with the input voltage V_{bc}. As frequency is varied, it is evident that the phase angle of V_f varies with respect to V_{bc}. This is shown in Fig. 6(B). Referring to Fig. 6(B)

$$V_f=V_{R4}-V_{R3} \tag{9}$$

$$V_{R4}=\frac{R_4}{R_4-jX_4}V_{bc}$$

$$V_{R3}=\frac{R_3}{R_3-jX_3}V_{bc}$$

$$V_f = \left(\frac{R_4}{R_4 - jX_4} - \frac{R_3}{R_3 - jX_3}\right) V_{bc}$$

In terms of R_3

$$V_f = \left(\frac{nR_3}{nR_3 - jR_{3/f}} - \frac{R_3}{R_3 - jnR_{3/f}}\right) V_{bc}$$

$$= \left(\frac{nf}{nf - j1} - \frac{f}{f - jn}\right) V_{bc}$$

$$= \frac{-jf(n^2-1)}{n(f^2-1) - jf(n^2+1)} V_{bc} \quad (10)$$

The value of the reference voltage V_r must be chosen such that at nominal frequency it equals V_f, since no compensating voltage is required at nominal frequency. Therefore

$$V_r = \frac{n^2-1}{n^2+1} V_{bc} \quad (11)$$

The frequency-compensating voltage then is

$$V_{comp} = V_r - V_f \quad (6)$$

$$= \left(\frac{n^2-1}{n^2+1} + \frac{j(n^2-1)f}{n(f^2-1) - j(n^2+1)f}\right) V_{bc} \quad (12)$$

In terms of positive- and negative-sequence component voltages this becomes

$$V_{comp} = j\sqrt{3}(V_1 - V_2) \times \left(\frac{n^2-1}{n^2+1} + \frac{j(n^2-1)f}{n(f^2-1) - j(n^2+1)f}\right) \quad (5)$$

Circuit Output

Having developed the voltage equations for both the filter circuit output and the compensating circuit output, these may now be added, as shown in Fig. 7. The output signal voltage may be stated simply as

$$V_{sig} = V_{sen} + V_{comp} \quad (7)$$

Effect of Load

To analyze further the circuit, the effect of loading the circuit may be considered. This can be done conveniently by considering the equivalent circuit shown in Fig. 10. Impedance of the isolating transformers is neglected. Here load current is

$$I_L = \frac{V_{sig}}{Z_T + Z_f + Z_s + Z_L}$$

and the load voltage is

$$V_L = \frac{Z_L V_{sig}}{Z_T + Z_f + Z_s + Z_L}$$

It can be shown that

$$Z_f = \frac{-n(n+1)R_3(1+jf)}{n(f^2-1) - j(n^2+1)f}$$

and

$$Z_s = \frac{-R_3(3+\sqrt{3})(1+jf)}{\sqrt{3}(f^2-1) - j4f}$$

where

$$n = \frac{R_4}{R_3}$$

From the foregoing material a circuit may be designed to operate with a particular load. It may be readily investigated over the frequency range to evaluate the error caused by frequency changes.

References

1. Symmetrical Components (book). C. F. Wagner, R. D. Evans. McGraw-Hill Book Company, Inc., New York, N. Y., 1933, chapt. 2.

2. Electrical Transmission and Distribution Reference Book (book). Westinghouse Electric Corporation, East Pittsburgh, Pa., 1950, chapt. 11.

Parallel Operation of Aircraft A-C Generators: Reactive Load Division and Selective Protection

H. A. KAHLE
ASSOCIATE MEMBER AIEE

THE parallel operation, as in today's aircraft, of electric a-c generators supplying a common bus requires that the total system load be divided equally among the paralleled generators. Real load and RL (reactive load) division are regulated through the speed and voltage regulator respectively. This paper concerns RL division only and the co-ordination of a high-phase voltage-sensing regulator with the RL sharing circuit and overvoltage and undervoltage protection.

Paper 56-435, recommended by the AIEE Air Transportation Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE South West District Meeting, Dallas, Tex., April 2-4, 1956. Manuscript submitted December 30, 1955; made available for printing February 24, 1956.

H. A. Kahle is with Jack & Heintz, Inc., Cleveland, Ohio.

The author wishes to thank Dr. H. M. McConnell for his support of the work and the many suggestive discussions on the subject.

Nomenclature

E_b = bus voltage
E_t = internal electromotive force of generator
E_1, E_2 = reference voltage of regulator
E_v = voltage at UVR (undervoltage relay) and OVR (overvoltage relay)
I = generator line current
I_c = current circulating among parallel generators
I_L = load current
x_d = synchronous machine reactance
x_{md} = generator mutual reactance, which combines E_t with generator field current
x = mutual reactance of UVR and OVR bias signal
$F(p)$ = amplifier transfer function
$G(p)$ = exciter transfer function
$H(p)$ = RL division circuit transfer function
S = sensitivity of RL division circuit
E_c = reference voltage
E_A = rectified bus voltage
E_B = additional direct voltage, e.g., from RL division circuit
R_i = resistance of first-stage magnetic-amplifier control winding
R_2, R_3 = load resistance of E_A and E_B respectively
R_4, R_5 = internal resistance of rectifier, filter, and transformer of E_A and E_B respectively
R_s = voltage divider for reference voltage
D_A, D_B = rectifier for bus voltage and RL division circuit respectively

RL Division

Present a-c generator RL division circuits consist of a current transformer connected to each generator, the secondaries connected in series, with additive polarity; see Fig. 1. Equalizing transformers ET connected to each current transformer apply a voltage to the bus voltage-sensing rectifiers which is nearly proportional to the difference between the line current of its corresponding generator and the average of all the generator line currents. Proper phase shift renders the rectifier output voltage insensitive to real load unbalance among the parallel generators. But, if a generator's reactive current is greater than the average value of the remaining generators, the output voltage of its rectifier increases and the generator's excitation is reduced by its regulator or vice versa.

A high-phase voltage-sensing regulator, Fig. 2, selects automatically the voltage of the highest phase and regulates it. Thus, the use of a RL division circuit, as described before, requires applying a signal corresponding to the RL unbalance to all

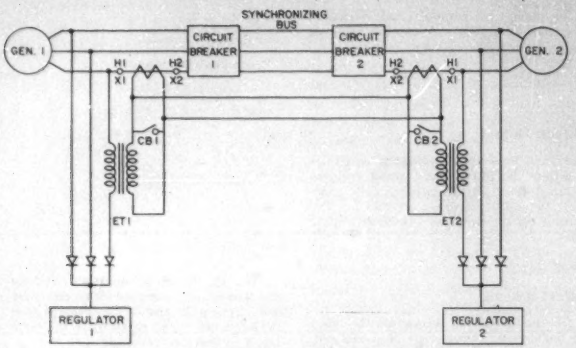


Fig. 1. RL division circuit for 2-parallel-generator system

three rectifier phase voltages, thus tripling the circuit of Fig. 1.

This, apparently, calls for applying the RL sharing signal to the regulator behind the bus voltage-sensing rectifier. At that point a direct voltage, its magnitude proportional to the generator RL unbalance and its polarity equal to the direction of the generator's reactive current, has to be introduced into the regulator error-sensing circuit. A circuit which gives the required properties is known as phase discriminator or ring demodulator and is shown in Fig. 3. A voltage E_2, opening and blocking the rectifiers alternatively each half-period, allows the voltage E_1 to drive a current through resistor R if the corresponding rectifiers are open. Thus, the magnitude of the d-c component of the current depends on the phase angle α between E_1 and E_2. A direct voltage appears across R

$$E = K_1 E_1 \cos \alpha; \quad (E_1 < E_2)$$

where K_1 is a constant.

Since the circuit has to give a direct voltage proportional to the reactive current unbalance in the parallel system, the desired signal is

$$E = K_2 I \sin \phi$$

where ϕ = load phase angle. Comparing both equations yields the desired voltage

$$E = K_1 E_1 \cos (\phi - 90°) = K_1 E_1 \sin \phi$$

by making the phase angle α equal to 90 degrees for $\phi = 0$. In a 3-phase system, this is easily accomplished by using the line-to-line voltage $T_2 - T_3$ opposite to the line-to-neutral voltage T_1; see Fig. 4. The voltage E_1 has to be proportional and in phase with the generator line current I in a 1-generator system and is taken from a resistive burden across the current transformer secondary. It is

$$E_1 = aIB$$

where

a = current transformer ratio
B = current transformer burden

The direct voltage across the resistor R

$$E = K_1 E_1 \sin \phi = K_1 BaI \sin \phi \qquad (1)$$

is therefore proportional to the reactive component of the line current in a 1-generator system, or proportional to the difference in reactive current between its corresponding generator and the average current of the other generators operating in parallel, as shown in Appendix I. The voltage E also changes polarity as the load power factor varies from lagging power factor ($\phi > 0$) to leading power factor ($\phi < 0$).

Equation 1 shows that the direct voltage is not affected by the line voltage E_2 as long as it is bigger than the voltage E_1. At the moment E_1 becomes larger than E_2, both voltages reverse function and voltage E_1 in turn switches the rectifiers. During this mode of operation, the direct voltage is proportional to the line voltage, which is almost constant. Thus, the d-c output voltage will remain constant and independent of the differential current in the current transformer loop. The desired knee in the droop characteristic,[1] with the use of the phase discriminator, does not depend on magnetic saturation in the transformer, but is established by the relative magnitude of the two voltages applied to the phase discriminator.

Regulator System

A complete regulator, incorporating a high-phase voltage-sensing circuit and the described RL division circuit, is shown in Fig. 2. It uses a glow tube as the reference voltage. The generator voltage is set by a potentiometer across the reference voltage tube. The high-phase voltage-sensing circuit rectifies each phase

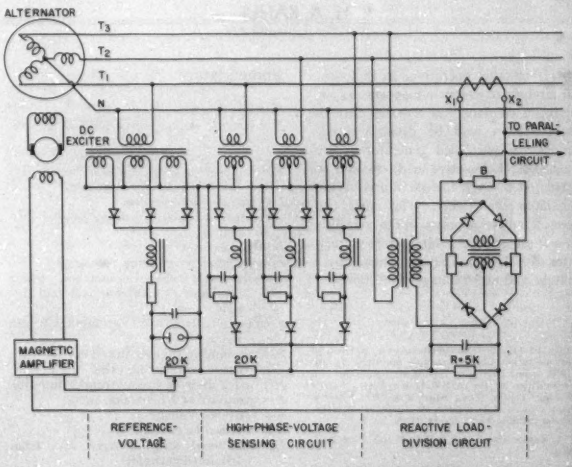


Fig. 2. High-phase regulator and RL division circuit

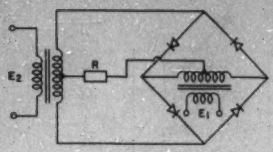


Fig. 3. Phase discriminator

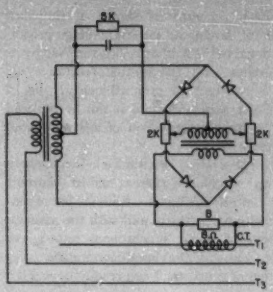


Fig. 4 (right). RL unbalance-sensing circuit

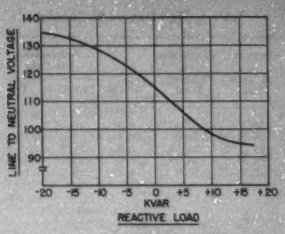


Fig. 5. Droop characteristic of one 40-kva generator

voltage individually. The three resulting direct voltages are applied to a second rectifier set at which the highest voltage blocks the other two rectifiers producing a direct voltage proportional to that a-c phase voltage the average value of which is highest. A 2-stage magnetic amplifier controls the field winding of a d-c exciter which, in turn, controls the synchronous generator.

The introduction of an additional 5,000-ohm resistor R into the control winding loop of the magnetic amplifier, as done here by connecting the RL division circuit in series with the control winding, might at first glance look like a considerable loss in gain. But, that is not necessarily true if, by proper circuit design, the internal impedance fo the RL division circuit, looking from R back into it, is kept low and so provides a by-pass for the magnetic-amplifier control current; see Appendix II. The internal impedance in the tested system is in the order of 600 ohms.

Test Results

A voltage droop characteristic taken on a single 40-kva generator with high-phase regulator and RL division circuit is shown in Fig. 5. It gives the generator voltage as a function of 3-phase RL. Note that the curve represents the voltage of the highest phase since the load-sharing signal is introduced into the regulator behind the phase-sensing element.

The sensitivity of RL division circuits is defined as the slope of this curve in the linear region[2] and can be expressed as generator voltage increment in volts per kilovar load of all three phases or, in a pu (per-unit) system, as

$$S = \frac{\text{generator voltage increment, pu}}{\text{lagging reactive current, pu}}$$

This corresponds also to the pu value of a virtual series impedance presented by the RL division circuit. As intuitive as this virtual impedance might be in understanding system operation, it loses its significance in the system analysis. The sensitivity of the present circuit is chosen to be 1.8 volt droop per lagging kilovar, equal to $S = -0.6$ pu.

The sensitivity S must have a certain minimum value to assure RL division among the paralleled generators within certain specified limits. If one generator's voltage differs from the average voltage of the remaining generators, then this generator will carry more or less than average RL according to the voltage droop characteristic. The generator voltage will finally settle at an intermediate point along this curve. At that point, a differential current, Appendix I, is maintained in the current transformer equalizer loop by an RL unbalance U which in pu values

$$U = \left| \frac{\Delta V}{2 I_R S \frac{n-1}{n}} \right|$$

where

ΔV = voltage difference of generator against average voltage of remaining generators, before paralleling, pu
I_R = reactive current, 0.66 pu at rated load
n = number of generators

Table I shows some results of measured unbalance on two parallel 40-kva generators. Test data are in good agreement with calculated data.

Co-ordination of Overvoltage and Undervoltage Protection Circuit with the Regulator

Overvoltages and undervoltages in aircraft a-c electric power systems can cause damage to electric equipment. The installation of OVR and UVR as part of protection circuits in single- or multigenerator systems is necessary since utilization equipment damage might result from such failure.

The OVR must have an inverse time characteristic, as specified in Military Specification *MIL-E-7894*[3], to avoid nuisance tripping during transients. In steady-state operation, as considered here, the trip points of the OVR and UVR are set at a voltage above and below rated terminal voltage. In a single-generator system, these points are fixed and the relays will trip when the generator voltage exceeds those limits, thus de-energizing the generator.

In a multigenerator system, the generators are connected to a bus. The bus will be kept at some intermediate voltage, nearly its rated value, even if a fault exists on one generator, resulting only in a large circulating current between generators. Thus, with many generators in parallel, the bus voltage is not likely to vary sufficiently to operate the OVR or UVR. If the bus varies sufficiently, all OVR or UVR are likely to trip, thus disconnecting all generators indiscriminately.

The phasor diagram of a 2-parallel generator system is shown in Fig. 6. Both generators are assumed to be of the round rotor type. The bus voltage E_e supplies a load current I_L to the load. If the excitations are different, the internal voltages E_{i1} and E_{i2} respectively are obtained from the bus voltage by adding to it the voltage drops across the synchronous reactances $I_1 x_{s1}$ and $I_2 x_{s2}$. This difference between E_{i1} and E_{i2} results in a circulating current I_c through both generators.

The current in each generator is

$$I_2 = 1/2\ I_L + I_c$$
$$I_1 = 1/2\ I_L - I_c \qquad (2)$$

and it is

$$E_{i2} - E_{i1} = E_e = I_2 x_{s2} - I_1 x_{s1} \qquad (3)$$

Inserting the generator currents yields

$$E_{i2} - E_{i1} = E_e = 1/2\ I_L x_{s2} + I_c x_{s2} - 1/2\ I_L x_{s1} + I_c x_{s1}$$

which gives the circulating current

Table I. Test Data on RL Sharing

Generator 1, Kilovar Total	Generator 2, Kilovar Total	Voltage	Unbalance, Per Cent Calculated	Unbalance, Per Cent Test
25.1	25.43	0.3	0.67	1.3
26.0	24.3	3.4	7.5	6.8
26.25	23.55	5.0	11.0	10.8
26.65	23.55	6.0	13.2	13.6
27.2	22.8	9.2	20.2	17.6

$$I_c = \frac{E_{t2} - E_{t1}}{x_{d1} + x_{d2}} + 1/2\, I_L \frac{x_{d1} - x_{d2}}{x_{d1} + x_{d2}} \qquad (4)$$

This current circulating between generators operating in parallel can readily be used to obtain selective overvoltage and undervoltage protection for the individual generators.

The current has a lagging power factor for overexcited generators and a leading power factor for underexcited generators. Applying a voltage to the OVR and UVR proportional in magnitude and phase to the circulating current results in selective protection.[4]

The voltage seen by the OVR and UVR of the overexcited generator 2 is

$$E_{r2} = E_b + x_d I_c \qquad (5A)$$

and of the underexcited generator 1

$$E_{r1} = E_b - x_d I_c \qquad (5B)$$

Equations 2 through 5 can be expressed in block diagram form, as shown in Fig. 7.

Each generator is shown in simplified form, having an internal voltage proportional to the field current i_f supplied by its exciter with the mutual reactance x_{md} as constant of proportionality. For a closer approximation, it can be replaced by the representation of a synchronous machine.[5]

The system is a multiple-loop system with circulating current and load current as additional inputs. With the use of conventional methods, and with the assumption of equal synchronous impedances $x_{d1} = x_{d2} = x_d$, the closed-loop transfer function of generator 1 can readily be written as

$$E_b = \frac{F(p)G(p)x_{md}}{1 + F(p)G(p)x_{md}} \times \left[E_1 - I_L \frac{x_d}{2F(p)G(p)x_{md}} + I_c \left(\frac{x_d}{F(p)G(p)x_{md}} - H(p) \right) \right] \qquad (6)$$

The sensitivity of the RL division circuit is defined as increment in generator voltage because of the circulating current. The response of bus voltage to circulating current is found from equation 6 by keeping E_1 and I_L fixed and it yields to

$$S = \frac{E_b}{I_c} = \frac{F(p)G(p)x_{md}}{1 + F(p)G(p)x_{md}} \times \left(\frac{x_d}{F(p)G(p)x_{md}} - H(p) \right) \qquad (7)$$

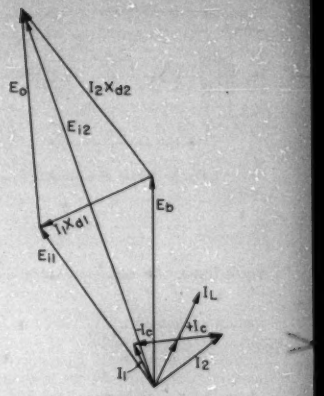


Fig. 6. Phasor diagram of two parallel generators

$$S \approx -\frac{F(p)G(p)x_{md}}{1 + F(p)G(p)x_{md}} H(p) \qquad (8)$$

since $x_d << F(p)G(p)x_{md}$.

The first part of equation 7 is caused by the voltage drop across the synchronous machine reactance, the second part is caused by the signal applied by the RL division circuit. Equation 7, which defines S in terms of system component

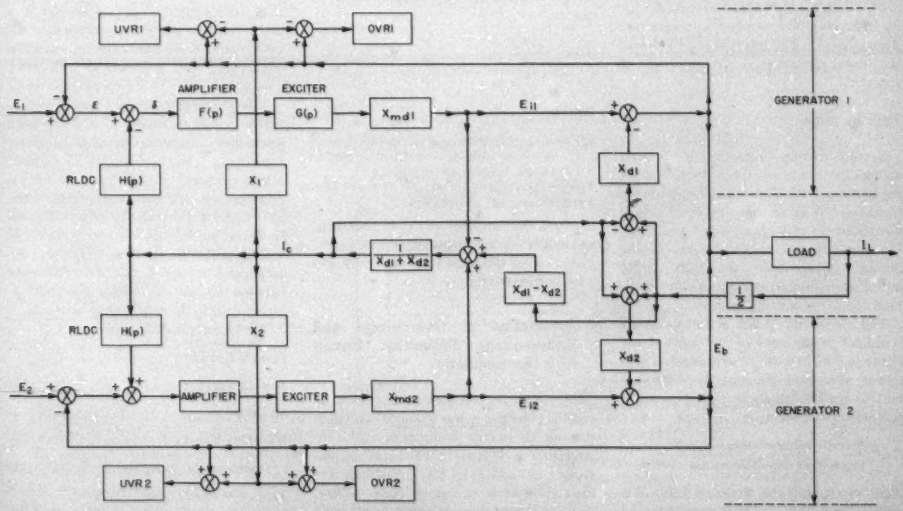


Fig. 7. Block diagram representation of two parallel generators

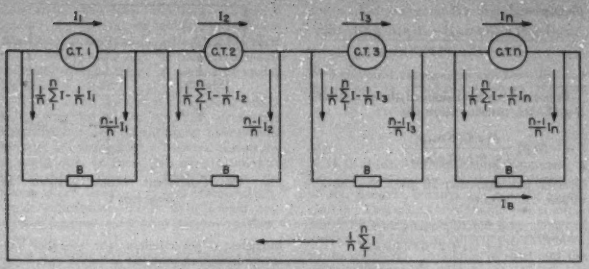


Fig. 8. Current transformer loop for parallel generator system

transfer functions, inserted into equation 6 gives the closed-loop transfer function for generator 1

$$E_B = \frac{F(p)G(p)x_{md}}{1+F(p)G(p)x_{md}} \times \left(E_1 - I_L \frac{x_d}{2F(p)G(p)x_{md}}\right) + SI_c \quad (9)$$

The bus voltage varies linear with circulating current. Comparing equation 9 with equation 5(B) demonstrates that

$$S = X \text{ (in steady state)}$$

avoids nuisance tripping of OVR and UVR. This comparison demonstrates that an excitation fault may be thought of as a change in the reference.

Steady-state operation requires the characteristics of the relays to be parallel to the RL division characteristic; see Fig. 5. Thus they have to have the same slope, and saturation of each circuit has to occur at the same current.

S and X are both functions of the complex frequency p. It is not necessary to have S equal X over the full frequency range; they have, however, to equal each other in the range of frequencies in which the relays are capable of responding.

Conclusions

The use of a phase discriminator as RL unbalance-sensing element in connection with a high-phase regulator has been described. The circuit is entirely insensitive to real load in a balanced system, and saturation of the generator voltage droop characteristic is established by the ratio of the two voltages applied to the phase discriminator rather than by magnetic saturation.

A block diagram of two synchronous generators operating in parallel is presented which has proved to be useful in system analysis and stability investigation. It is shown that UVR and OVR have to be biased by a voltage proportional to differential reactive current, yielding in equal slope of the generator voltage droop and relay characteristics to avoid nuisance tripping of the OVR and UVR during steady-state and transient reactive load unbalance. The examination of RL division stability restricts the maximum permissible sensitivity of the RL division circuit to the value of the unsaturated synchronous reactance of the generator.

Appendix I. Analysis of RL Division Circuit

The secondary windings of the current transformers of n parallel generators are connected in series with additive polarities; see Fig. 8. The burden B of each current transformer corresponds to the resistor B of Fig. 2. For calculation of the circuit sensitivity, an expression for the differential current I_B through B in terms of current transformer secondary currents I will be derived, with the assumption of ideal current transformers.

Each current transformer is loaded with a network consisting of the parallel circuit of one resistor B and $(n-1)B$. The current transformer secondary current I_n of the nth generator splits into two components, $I_n(n-1)/n$ through its corresponding current transformer burden and I_n/n through the remaining current transformer burdens.

The differential current in the nth burden is

$$I_B = \frac{1}{n}\sum_{i=1}^{n} I_i - \frac{1}{n} I_n - \frac{n-1}{n} I_n$$

$$= \frac{1}{n}\sum_{i=1}^{n} I_i - I_n \quad (10)$$

where $I_1, I_n \ldots I_n$ is the current transformer secondary current of the 1st, 2nd, . . . nth generator respectively.

Defining the average current of the remaining $(n-1)$ generators as

$$I_{av_2} = \frac{1}{n-1}\sum_{i=1}^{n-1} I_i \quad (11)$$

Equation 10 becomes

$$I_B = \frac{1}{n}\sum_{i=1}^{n-1} I_i + \frac{1}{n} I_n - I_n$$

$$= \frac{1}{n}\sum_{i=1}^{n-1} I_i - \frac{n-1}{n} I_n$$

$$I_B = \frac{n-1}{n}(I_{av_2} - I_n) \quad (12)$$

The differential current sensed by the RL division circuit is equal to the difference of the respective generator's line current and the average line current of the remaining generators. Its amount varies with the number of generators operating in parallel.

Appendix II. Analysis of Regulator Error-Sensing Circuit Activity

Fig. 9 shows schematically the sensing circuit of the magnetic amplifier used in the system. Evaluating the circuit, an equivalent circuit, Fig. 10, can be drawn which gives the current I in the magnetic-amplifier control winding as

$$I = \frac{E}{R}$$

where

$$E = \alpha E_C - E_A \frac{R_3}{R_3+R_4} - E_B \frac{R_2}{R_2+R_5}$$

$$R = R_1 + \alpha R_3 + \frac{R_3 R_4}{R_3+R_4} + \frac{R_2 R_5}{R_2+R_5}$$

provided that the current driven by E_A and E_B over their respective rectifiers D_A and D_B is bigger than the current I, thus providing a by-pass for the magnetic-amplifier control current over the internal voltage source resistance. The foregoing equation indicates that the sensitivity of the circuit

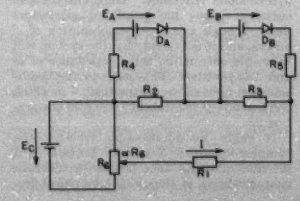


Fig. 9 (left). Error-sensing circuit

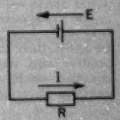


Fig. 10 (above). Equivalent circuit of Fig. 9

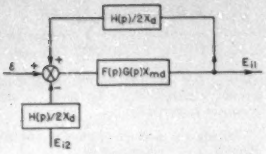


Fig. 11. Minor loop of RL division circuit

will not be decreased, if, for any additional voltage E_B, the resistance

$$\frac{R_3R_4}{R_3+R_4} << R_1+\alpha R_4+\frac{R_2R_4}{R_2+R_4}$$

R_3 can be high as long as the internal voltage source resistance R_4 is small enough.

Appendix III. Stability of Parallel Generators

A parallel generator system becomes a multiple-loop system, as shown in Fig. 7. A regenerative feedback exists around the minor loop, formed by $F(p)$, $G(p)$, x_{md}, and $H(p)$. Its closed-loop transfer function, Fig. 11, for generator 1 is equal to

$$E_{t1}=[\epsilon-E_{t2}H(p)/2x_d]\times\frac{F(p)G(p)x_{md}}{1-H(p)F(p)G(p)x_{md}/2x_d}$$

$$E_{t1}=[\epsilon-E_{t2}H(p)/2x_d]Q(p) \quad (13)$$

Replacing the minor loop by its closed loop transfer function results in the simplified block diagram of Fig. 12.

To demonstrate the effect of RL division sensitivity on the over-all system stability, the closed-loop transfer function of Fig. 12 has to be investigated. The system has three inputs

E_i

$E_{t2}H(p)/2x_d$

$I_ix_d=1/2\ I_Lx_d-I_cx_d$

The last term can be thought as acting at the input by dividing it by $Q(p)$

$$I_ix_d/Q(p)=1/2\ I_Lx_d/Q(p)-I_cx_d/Q(p)$$

The closed-loop transfer function becomes then

$$E_b=\frac{Q(p)}{1+Q(p)}\,[E_i-E_{t2}H(p)/2x_d+I_cx_d/Q(p)-I_Lx_d/2Q(p)] \quad (14)$$

The characteristic equation of the system is given by $1+Q(p)=0$. For the system to be stable, there must not be any roots with positive real parts of the characteristic equation which, with the use of $Q(p)$ from equation 13, results in

$$1+Q(p)=1+\frac{F(p)G(p)x_{md}}{1-H(p)F(p)G(p)x_{md}/2x_d}=0 \quad (15)$$

Clearing fraction

$$1+Q(p)=\frac{2x_d[1+F(p)G(p)x_{md}]-H(p)F(p)G(p)x_{md}}{2x_d-H(p)F(p)G(p)x_{md}}=0 \quad (16)$$

Equation 8 which defined the sensitivity S of the RL division circuit as

$$S\approx -H(p)\,\frac{F(p)G(p)x_{md}}{1+F(p)G(p)x_{md}} \quad (8)$$

inserted into equation 16 gives the characteristic equation to

$$1+Q(p)=\frac{[1+F(p)G(p)x_{md}]\,[2x_d+S]}{2x_d+S[1+F(p)G(p)x_{md}]}=0$$

or

$$[1+F(p)G(p)x_{md}]\,[2x_d+S]=0 \quad (17)$$

The closed-loop transfer function in terms of load and circulating current is given by equation 6. If the generators are not paralleled and the RL division circuit is not effective, there will be no circulating current. The single-system closed-loop transfer function

$$E_b=\frac{F(p)G(p)x_{md}}{1+F(p)G(p)x_{md}}E_i-1/2\ I_L\times\frac{x_d}{1+F(p)G(p)x_{md}} \quad (18)$$

is assumed to be stable; thus, its characteristic equation

$$1+F(p)G(p)x_{md}=0 \quad (19)$$

will not have any roots with positive real parts. The characteristic equation to be investigated, equation 17, shrinks therefore to

$$2x_d+S=0 \quad (20)$$

The sensitivity of the RL division circuit is negative according to definition. It follows that

$$-S=2x_d \quad (21)$$

is the limit for parallel-type stability. The RL division sensitivity of a 2-generator system is one-half of the single generator's sensitivity; see Appendix I. The general stability criterion for n parallel generators can therefore be stated as

$$-\frac{n-1}{n}\,S<x_d \quad (22)$$

It has been shown that the sensitivity of the RL division circuit may also be defined as the pu value of a virtual series impedance which, according to the definition of S, is a negative impedance. This negative impedance can be thought as being in series with the internal generator impedance, the synchronous reactance. Only current circulating among the paralleled generators flows through this series combination of S and x_d, and it is evident that the system will be stable only if the total impedance of x_d and S is positive. This leads to the condition that x_d has to be bigger than $-S$, which corresponds to the stability criterion of equation 22.

It might be questionable whether the complex quantity S can be limited to a machine quantity since both have different characteristics in the range of critical frequencies. Nevertheless, the stability criterion indicates that the absolute values for S and x_d must stay in the given limits of equation 22 because the load disturbance transfer functions of S and x_d are of the proportional type and can be approximated by their absolute values at low frequencies.

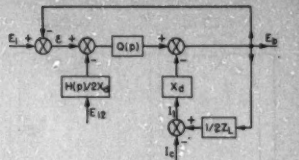


Fig. 12. Equivalent block diagram of generator 1

References

1. Final Report on Aircraft A-C Power Systems, Part 5, E. S. Sherrard. *Memorandum Report 185*, Naval Research Laboratory, Washington, D. C., 1953.

2. Sensitivity Requirements of Reactive Load Division Circuits in Aircraft Electric Systems, Emile S. Sherrard. *AIEE Transactions*, vol. 73, pt. II, Nov. 1954, pp. 279–88.

3. Aircraft Characteristics of Electric Power. *Military Specification MIL-E-7894*. Washington, D.C., Aug. 14, 1952.

4. Overvoltage Detection in Single and Multigenerator Aircraft A-C Systems, W. M. Tucker, M. Trbovich. *AIEE Transactions*, vol. 73, pt. II, Nov. 1954 (Jan. 1955 section), pp. 420–25.

5. Study of Voltage Regulating Systems for Aircraft Alternating Current Generators. Robert M. Saunders, David C. White. *WADC Technical Report 55–311*, Wright Air Development Center, Dayton, Ohio.

Discussion

S. C. Caldwell and R. T. Smith (General Electric Company, Schenectady, N. Y.): Mr. Kahle has presented a relatively simple method of obtaining control on the shape of reactive division sensitivity curves. This shape is of importance in considering parallel-system operation when a wire or a current transformer has opened in the differential current transformer loop supplying the RL division circuitry of each regulator in the system. Usually, high sensitivity is desired to obtain good reactive division, in spite of differences in regulator operating points, and to permit the use of sizable signals for protection against malfunctions in the excitation systems. Then, if the droop curve is linear, system voltage will sag to a low value but reactive division will be maintained. If sharp saturation characteristics are introduced to restrict the voltage drop under these conditions, then reactive division is obtained under open-loop conditions only up to the saturation point. RL in excess of this quantity will not be divided among the machines.

With the circuitry presented in this paper, it appears feasible to effect a compromise where voltage sag is restricted but reactive division is maintained to a reasonable de-

gree under open-loop conditions. Thus, a system can be evolved which can continue operating indefinitely without (in most cases) overloading any of the machines or reducing voltage below a value still usable for many loads.

We have felt that, in most applications, it is preferable to incorporate protection against open-loop operation so that normal voltage is restored to all source busses by opening one or more tie circuit breakers under this condition. If this is done, the shape of the droop curve is of comparatively small importance except as it affects protection against excitation system malfunctions.

As a matter of interest, we think this reactive division circuitry will exhibit some sensitivity to real power once a reactive unbalance exists when the level of the reference voltage is made small enough to limit the droop, as shown in Fig. 5. If we understand the operation of the circuitry correctly, this will occur even in the linear range. Some recent work[1] has shown that system stability is sometimes greatly affected by real load sensitivity in the reactive division loop. Since it may be expected that some reactive unbalance will exist in normal operation and large amounts may exist during system malfunction conditions, it would appear useful to the system designer and to those conducting stability analyses to have some indication of the degree of real load sensitivity obtained under both normal and abnormal RL unbalance conditions.

Reference

1. Frequency Modulation and Load-Division Instability in 400-Cycle Aircraft Electric Systems, Henry Oman. *AIEE Transactions*, vol. 75, pt. II, 1956 (*Paper 56-436*).

H. A. Kahle: The supporting comments of Mr. Caldwell and Dr. Smith on malfunction of RL division circuits are greatly appreciated. The selection of the proper point at which the characteristic starts saturating will always be a compromise between the extent up to which RL division is desired and the permissible voltage sag.

The expression for maximum permissible sensitivity of the RL division circuit has been, to the author's knowledge, derived first by Sherrard (see references 1 and 2 of the paper) for an infinite number of generators in parallel by expressing the sensitivity in terms of several partial differential quotients. With the use of servomechanism analysis, the block diagram lent itself to the expression of maximum permissible sensitivity for any number of generators operating in parallel.

Raising the question of real load sensitivity, it seems to be wise to define real load first. Since the circuit, as presented, senses one phase only, all consideration should be restricted to a balanced system, real load then being defined as $3\,EI\cos\phi$, where ϕ is the phase angle between line current and line-to-neutral voltage in the balanced system. The output voltage of the phase discriminator will be a maximum if the phasors of the current in phase T_1 and the voltage $T_2 - T_3$ are parallel, as for reactive load, and will be zero if both phasors are perpendicular to each other, as for real load. This holds during any mode of operation of the phase demodulator, even when the level of the reference voltage is small. The circuit will therefore never exhibit an output to real load as long as the system is balanced. This has been verified by tests on a single generator in which the voltage was not affected from no load to rated real load.

Deliberate real load sensitivity in the RL division loop has been discussed in the past as a means to effect system stability. I feel a statement in this general form should be made cautiously since it could not be verified by tests on one specific system. The parallel system is a multiple-loop system from the standpoint of servomechanisms, in which the component transfer properties can be completely different for individual systems. Thus, changing components might or might not affect the system performance, depending only on the transfer functions of the components and the system. Dynamic system analysis will assist the engineer in designing systems and will probably be the only way to give an insight into system behavior.

Frequency Modulation and Load-Division Instability in 400-Cycle Aircraft Electric Systems

HENRY OMAN
MEMBER AIEE

THE OUTPUT frequency of an alternator is proportional to the rotating speed of the field flux of the alternator. Variations in this field flux speed can be caused by torsional flexing of the alternator shaft and by instability in the speed of the prime mover.

Paper 56-436, recommended by the AIEE Air Transportation Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE South West District Meeting, Dallas, Tex., April 2-4, 1956. Manuscript submitted July 13, 1955; made available for printing February 14, 1956.

Henry Oman is with Boeing Airplane Company, Seattle, Wash.

The author takes this opportunity to thank Walter P. Koenig, Jr., of the Boeing Airplane Company, for the use of circuits which he had developed for measuring frequency modulation and load oscillation.

When the output frequency of an aircraft alternator varies about a nominal value, frequency modulation can be said to be present. If these variations are cyclic, they may affect adversely the performance of sensitive equipment such as autopilots, fire control, and navigational computers.

Furthermore, unstable load division can result if one of paralleled alternators has frequency modulation. Other factors also may contribute to unstable load division. Cyclic changes in load division can be great enough to produce cyclic variations in the voltage of paralleled alternators. Such cyclic voltage variations can affect adversely sensitive equipment such as that just mentioned.

The amount of frequency modulation and load-division instability that can be present in an airplane electric system must be controlled. Satisfactory airplane performance in this respect can be attained only if the limits of permissible frequency modulation are defined in terms which cannot possibly be misinterpreted by the prime-mover manufacturer, the alternator and alternator control manufacturer, and the power utilization equipment manufacturer. An objective of this paper is to suggest ways of defining frequency and load-division instabilities.

Equipment Involved in Frequency and Load-Division Instability

Equipment in a 400-cycle electric power system which can be involved in frequency modulation includes prime movers, alternators, voltage regulators, load division controllers, and protective equipment.

The rotating speed of an aircraft engine, particularly a turbojet engine, usually can be varied by the pilot to give different amounts of thrust. Hence, in most modern airplanes 400-cycle alternators are driven by either of the following:

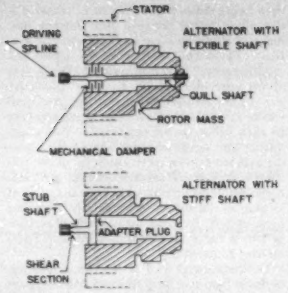


Fig. 1. Examples of shafts used in 400-cycle aircraft alternators

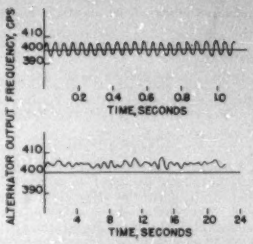


Fig. 2. Two examples of frequency modulation in output of 400-cycle aircraft alternator

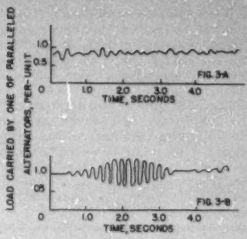


Fig. 3. Two examples of load-division instability. Power output of one of two paralleled alternators shown. Total system load was constant

1. A hydromechanical constant-speed drive[1,2] driven by a shaft from the engine. This device has an integral variable-displacement hydraulic pump and a constant-displacement hydraulic motor which make up the differential between variable input speed and a constant output speed. The output speed is controlled by a governor, which varies the displacement of the pump by means of a hydraulic piston and linkage.

2. An air turbine drive which runs on compressed air bled from the jet engines. The speed of the air turbine is controlled by a governor that varies the inlet air valve position by means of a suitable hydraulic or pneumatic piston and linkage.

Many aircraft alternators are equipped with a quill shaft and mechanical damper assembly as shown in Fig. 1. Such construction was found necessary for protecting generators coupled directly to reciprocating engines from shaft breakages as a result of engine torsional vibration and backfiring.

Causes of Frequency Modulation

Alternator Shaft

The quill shaft and alternator rotor assembly shown in Fig. 1 have a natural torsional-vibration frequency which is established by the spring constant of the shaft and the mass of the rotor assembly. This natural frequency might be in the order of 15 to 25 cps (cycles per second). If the prime-mover governor has a natural frequency which is about the same as the natural frequency of the alternator quill shaft and rotor assembly, then frequency modulation may be encountered. Some mechanical dampers, because of backlash, do not seem to be effective in preventing frequency modulation that originates from this source. An undamped quill shaft and rotor assembly may amplify torsional pulsations coming from the prime mover.

It is interesting to note the small torsional deflections in the shaft that will produce appreciable frequency modulation. For example, assume that the torsional vibration has an amplitude of $\theta_{max} = \pm 2$ degrees, and its rate ω is 18 cps. Then

$$\omega = (2\pi)(18) = 36\pi \frac{\text{radians}}{\text{second}}$$

$$\theta = (2)\left(\frac{\pi}{180}\right) \sin \omega t \text{ radians}$$

$$\frac{d\theta}{dt} = \frac{\pi}{90} \omega \cos \omega t$$

At $t = 0$ and $t = 1/36$ seconds, $d\theta/dt$ will be maximum.

$$\frac{d\theta}{dt_{max}} = \left(\frac{\pi}{90}\right)(36\pi)(\pm 1) = \pm 1.25\pi \frac{\text{radians}}{\text{second}}$$

With an 8-pole 6,000-rpm alternator, 400 cycles represents

$$\frac{d\theta}{dt} = \frac{(6{,}000)(2\pi)}{60} = 200\pi \frac{\text{radians}}{\text{second}}$$

Thus, this ±2-degree torsional vibration corresponds to a ±2.5-cps frequency modulation.

Frequency modulation resulting from alternator shaft resonances has been reduced by using a governor that has no resonances at the alternator shaft resonant frequency, or by using a stiff-shaft alternator such as is shown in Fig. 1. Such a stiff-shaft alternator may have a torsional resonant frequency which is three to five times that of a quill-shaft alternator.

Prime-Mover Governor

A governor which causes the prime mover to hunt in speed will produce frequency modulation. Most governors used in alternator drives operate isochronously; and speed droop, if needed for paralleling, is produced by biasing the governor with a signal derived from a load-sensing device. Such a governor tends to be unstable, and hence may produce frequency modulation. Frequency modulation originating from governor instability may be reduced by utilizing effective stabilization in the governor.

Some fairly good governors from a standpoint of frequency modulation have been produced. The most successful have electric discriminators that establish a frequency reference which is compared with the output frequency of a tachometer generator or the alternator. An electric governor of this type can be stabilized effectively by applying electric feedback signals into appropriate networks.

Effects of Frequency Modulation

Variations of airplane frequency about a nominal value can be cyclic or random. Often random variations have cyclic components. Typical frequency variations are shown in Fig. 2.

In general, motors are not seriously affected by small variations in frequency. Lamps and heaters are unaffected. Certain types of electronic equipment seem to be sensitive to changes in frequency, especially if such changes are cyclic. For example, an autopilot may be designed to have response rates of over 6 cps in order to move the control surfaces fast enough to limit airplane attitude deviations. Small changes in airplane electric power frequency occurring at a cyclic rate in the order of 6 cps can cause annoying "buzz" in the pilot's rudder pedals. Larger amounts of frequency modulation can cause movements in the airplane control surfaces. Radar-type equipment also seems to be sensitive to frequency modulation. One common effect is unstable or fuzzy images on cathode-ray tubes.

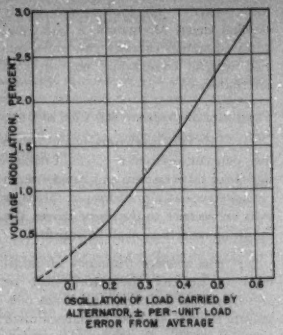


Fig. 4. Voltage modulation resulting from load division oscillation

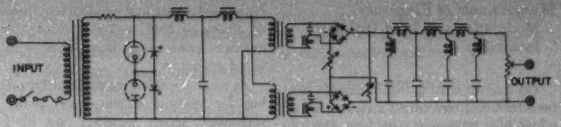


Fig. 5. Discriminator circuit for recording frequency modulation in 400-cycle voltage

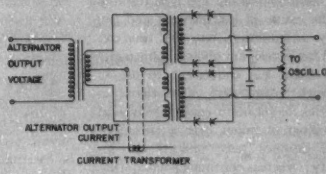


Fig. 6. Discriminator circuit for recording load division instability

In general, cyclic changes in frequency are more serious than random changes in frequency. Occasional frequency dips and rises resulting from load changes are expected in an airplane, and apparently have no serious or persisting effect on the electronic equipment.

Causes of Load-Division Instability Among Paralleled Alternators

The load carried by one of paralleled alternators is a function of its input torque. Any pulsations in input torque will affect the proportion of load a given alternator is carrying, with relation to the total system load. During load-division instability, the total system load will remain essentially constant unless the instability is so violent that system voltage is affected. Some examples of load-division instability are shown in Fig. 3.

The prime mover can be a source of pulsations in the input torque of an alternator. An unstable governor obviously will result in cyclically or sporadically varying torque and hence unstable load division. Changes in the input speed or inlet air pressure supplied to the prime mover can cause pulsations in the input torque to an alternator, especially if these changes occur so rapidly that the prime-mover governor cannot make compensating corrections.

A less obvious source of alternator input torque pulsations is the reactive load division network. The synchronous torque coefficient of an alternator is a function of many factors, including the amount of excitation. With high kva load and low-lagging power factor, the synchronous torque coefficient is high. With the same high kva load and unity power factor this coefficient is lower. If an alternator is carrying a constant real load, and the excitation is changed because of a signal which the voltage regulator receives from the reactive-current equalization loop, there must be a change in the rotor power angle. This power angle change represents a physical movement of the rotor, which in turn involves torques and inertias. Because the electrical spring between the inertias of paralleled alternators and prime movers has little damping, it is possible for substantial load oscillations to be built up if the spring constants and voltage regulator natural frequencies are such that resonance results. This type of load oscillation can be reduced by providing suitable damping in the system. Such damping may be in the prime-mover governor, the voltage regulator, or in the alternator amortisseur windings.

In one case, rather violent load division oscillation was stopped by changing the sensitivity of the voltage regulator to real power unbalance. In the original voltage regulator design a given amount of real current unbalance, as sensed through the equalizer loop, would produce only about 8% (per cent) of the effect on the voltage regulator which would be produced by the same amount of reactive current unbalance. The voltage regulator that stopped the load oscillation had been altered so that its response to real current unbalance was the same as its response to reactive current unbalance.

Another factor that affects load oscillation is the spring constant of the alternator shaft. In systems having a high-speed air turbine geared to a low-speed alternator, the turbine wheel, because of its high speed, possesses most of the inertia of the combination. In flexible shaft alternators, the shaft connecting the alternator rotor mass to the turbine wheel mass may have a spring constant which is in the same order as the synchronous torque coefficient of the alternator. In some cases, the susceptibility of a system to load oscillation can be reduced by substituting a stiff-shaft in place of the flexible shaft in the alternator.

Effects of Load-Division Instability

A small amount of load-division instability seems to be tolerable. With higher amounts of load division oscillation, undesirable effects are produced. The most serious is probably the resulting voltage modulation, which, if large enough, will affect certain sensitive electronic equipment. The relationship of load oscillation to voltage modulation for one particular system is shown in Fig. 4.

The voltage modulation allowed by military specifications on aircraft 400-cycle 120/208-volt systems is generally 2%. However, some electronic equipment has been found to deteriorate in performance when the voltage modulation is even less than 2%.

The voltage modulation frequency observed in oscillograms was twice the load oscillation frequency. One example of load oscillation was observed to have a cyclic rate of 5 cps. Thus, the resulting voltage modulation would be at a cyclic rate of 10 cps, which is a rate that would affect the autopilot. One particular autopilot configuration has been observed to be more sensitive to voltage modulation than to frequency modulation.

Measurement of Frequency Modulation and Load-Division Instability

The direct-writing oscillograph has been the most convenient instrument for measuring frequency and load-division instability. These instruments have response rates up to 100 cps, which seems to be adequate for these phenomena. Light beam oscillographs are useful in research into the causes of load-division and frequency instabilities, but the time required for developing photographic paper makes them less convenient to use in this work. Also, very often these instabilities do not occur continuously, and an oscillogram taken over a period of several seconds may indicate no instability; whereas continuous monitoring with a direct-writing oscillograph will indicate sporadic instability.

The circuits used to convert frequency-stability and load-division information into oscillograph traces have been the subject of much misunderstanding. Often, actual instabilities in the prime mover have been blamed on the instrumentation. At other times, the instrumentation has not had a sufficiently high frequency response to reveal the instabilities that exist. Some instrumentation has had such a low response that it was difficult to differentiate between frequency instability and instrumentation noise.

A circuit that has produced satisfactory oscillograms of frequency instability is shown in Fig. 5. This is basically a frequency discriminator circuit and has accurate response to frequency modulation having cyclic rates from 0 to 125 cps.

A circuit which has been found suitable for recording load-division instability having cyclic rates up to 60 cps is shown in Fig. 6. This circuit contains a discriminator that produces a signal which is proportional to the real-power component of the alternator output current.

Suggested Specification for Frequency and Load-Division Stability

Obviously, limits must be established for frequency and load-division stability in the a-c electric power system of any airplane. Previous attempts at specifying these stability limits have resulted in difficulties because the same interpretations were not placed on the specifications by all of the various persons involved.

Stability limits must be specified for both steady-state and variable conditions of alternator load, and for all possible values of prime-mover input speed or inlet air energy.

To make the following suggested wordings for limits of stability more understandable, actual numbers are used. These numbers represent limits which have been found workable on large airplanes equipped with 400-cycle a-c power systems.

Suggested Definitions

Average Frequency

Average alternator output frequency is the average value of the instantaneous alternator output frequency during a one-minute period. Average frequency may be determined with an instrument which continuously records instantaneous frequency for the one-minute period.

Average Load-Division Error

Average load-division error between two paralleled alternators is the average of the instantaneous load of one alternator during a one-minute period minus the average of the instantaneous load of the other alternator during this one-minute period. If more than two alternators are paralleled, the load-division error shall be measured on the two alternators that result in the largest load division error.

Instantaneous Frequency

Instantaneous alternator output frequency shall be that value of frequency which is proportional to the angular velocity of the alternator field structure at any instant. (Note: Changes in power factor and harmonic content can also momentarily affect alternator output frequency. In practice, these effects are small.)

Instantaneous Load-Division Error

Instantaneous load-division error among paralleled alternators shall be the difference between the largest and smallest kilowatt load carried by alternators operating in parallel at any one instant.

Per-Unit Load

The rated load of the system or of one alternator, whichever is being considered, is 1.0 per-unit load.

Rated System Capacity

A load equal to the rating of one alternator times the number of alternators in the system being considered is rated system capacity. (If load is not evenly divided among alternators, a load equal to rated system capacity cannot be carried without overloading one or more alternators.)

Frequency Stability, 400-CPS System

1. With the alternators connected to any constant-resistance constant-inductance load between zero and rated system capacity, and at all rated input speeds or inlet air energies to the prime mover, the following conditions shall be satisfied.

(*a*) The average frequency of paralleled or isolated alternators shall be between 396 and 404 cps, and in any one-minute period it shall not deviate by more than 1.0 cps from the average output frequency in any other one-minute period during one hour of continuous operation.

(*b*) The instantaneous output frequency of isolated or paralleled alternators during any one-minute period shall not deviate from the average frequency in that one-minute period by more than 1.0 cps.

Note: If load division in a paralleled system is achieved by having the prime-mover speed droop as load is increased, then frequency stability will have to be specified in terms of allowable deviation from a specified speed-load curve.

2. With any constant prime-mover input speed or inlet air energy and with paralleled alternators, sudden application or removal of any load up to rated system capacity shall not cause the instantaneous alternator output frequency to deviate more than ±20 cps from the average frequency of the one-minute period starting 1.0 second after the load change.

3. With any constant prime-mover input speed or inlet air energy and with paralleled alternators, the alternator output frequency shall return to steady-state limits within 1.0 second after applications of a 1.0 per-unit system shock load, and within 0.5 second after application of a 0.5 per-unit shock load.

4. With any constant input speed or inlet air energy to the prime movers, and with paralleled alternators carrying constant overload, the instantaneous frequency shall be within the following limits:

Permissible Instantaneous Output Frequency Deviation

Alternator Load Per Unit	Per Cent of Deviation Allowed for 100% Load
1.5	150
2.0	200

LOAD-DIVISION STABILITY

1. The following requirements shall be met at all times in a system containing paralleled alternators carrying any load between zero and rated system capacity, and with any constant input speed or inlet air energy being supplied to the prime movers.

(*a*) The average load division error shall not exceed 10% of the kilowatt rating of one alternator.

(*b*) The instantaneous load division error shall not exceed 20% of the kilowatt rating of one alternator.

2. With paralleled alternators driven by prime movers being supplied with any constant input speed or inlet air energy, sudden application or removal of any load up to rated system capacity shall not cause an instantaneous load division error of greater than 40% of the kilowatt rating of one alternator.

3. These load division requirements shall be applicable regardless of the number of alternators paralleled on the plane.

Conclusion

The numerical values for stability limits used in this paper were found to be satisfactory for one airplane. Each new airplane electric system has to be analyzed to find the maximum possible stability limits that the power-consuming components can withstand. Obviously, wider stability limits usually result in less cost and weight in the prime-mover speed control.

The intent of this paper is not to suggest numerical stability limits, but rather to suggest a way of specifying stability so limits that they will be:

1. Adequately comprehensive.

2. Absolutely clear to the manufacturers of prime movers and power utilization equipment.

It is hoped that as a result of discussion initiated by this paper, a standard way of specifying frequency and load division stability will evolve.

References

1. PERFORMANCE OF A CONSTANT-SPEED DRIVE, E. W. Gilboy. *AIEE Transactions*, vol. 73, pt. II, Sept. 1954, pp. 179–84.

2. NEW DEVELOPMENTS IN DIFFERENTIAL-TYPE HYDRAULIC TRANSMISSIONS AND CONTROLS, R. H. Eisengrein. *Ibid.*, vol. 72, pt. II, 1953 (Jan. 1954 section), pp. 423–29.

Discussion

H. M. McConnell (Jack & Heintz, Inc., Cleveland, Ohio): This paper treats a problem which is becoming increasingly important. It should be emphasized that the two primary causes for load transfer oscillations in a parallel a-c system are excitation modulation and torque modulation. The first will produce reactive load transfers and the second will produce real load transfers. If torque modulation is present it will appear in an isolated system as a frequency modulation effect; however, if two generators in parallel operate with reinforced torque modulations the system frequency will remain relatively unchanged and load transfer will take place.

The effect on torque modulations produced by generator quill shaft flexibility must be related to the output inertia of the drive. For example, the author mentions the fact that an air turbine drive has quite high output inertia, and consequently a resonant condition may occur as set by the generator inertia and the quill shaft flexibility. On the other hand the output inertia of hydraulic constant speed drives is negligibly small compared to the generator inertia, and the quill shaft flexibility is accordingly unimportant in investigating load transfer problems.

The drive and generator are each closed loop systems having several minor loops, interconnected by real and reactive load division circuits so that the combined performance of the entire system must be considered. Therefore, any results that have so far appeared bearing on this problem must be evaluated considering that they apply to the particular system on which the investigation was made, and do not necessarily have general significance.

The stability in such a system depends on the dynamic characteristics of the components, some of which vary widely. Generators will not be significantly different in per-unit transient response since they are designed to common specifications of power capability and overload. However, exciter and regulator dynamics can vary remarkably from one system to another. Also, drives of similar construction will have similar dynamic performance in the basic power portion, but the governor and trim dynamics can be made to vary widely. These remarks lead to the conclusion that the over-all stability problem can be approached intelligently only on a system basis.

One simple situation can be considered which will illustrate the way in which generators affect this problem. A simplified analysis will be presented to show how cyclic load transfer between machines can take place, while average real and reactive loads are divided equally in the normal manner. Fig. 7 shows the vector diagram of a salient pole a-c generator-supplying load of power factor, $pf = \cos\theta$. The steady-state portion of this diagram holds for each of the two machines, since on the average the real and reactive loads are equally divided due to the regulating action of the two controllers. It is assumed that any hunting of one machine against the other results in electrical oscillations superposed upon the average state. This assumption neglects variations in terminal voltages due to the hunting and consequently neglects any damping effect that the load might have on the oscillations.

Fig. 7 shows also a small angular displacement δ' appearing between the two internal voltages E_I and E_{II}. Since the electric connection between machines is relatively stiff, consisting of a transmission line with no more than a few per cent reactance, the circulating current is limited essentially by the transient reactances of the machines and is proportional to the vector difference between the voltages behind transient reactances E_I' and E_{II}'.

The circulating current is reactive with respect to $\Delta E' = E_I' - E_{II}'$, and is by symmetry real with respect to the internal voltage E of Fig. 7. The power factor of the circulating current with respect to the terminal voltage becomes $\pm\cos\delta$, and the reactive factor is $\pm\sin\delta$. Therefore the effect of this perturbation respectively upon the real and reactive load division servos is dependent upon the average load. This fact must be taken into account when interpreting the performance of these servos under the conditions discussed here.

The electrical stiffness of the perturbed system can be computed simply from a knowledge of the voltage behind transient reactance, which is again a function of average load. For two equal machines without shunt loading (the situation here in view of the basic assumption discussed earlier) the swing equation for small angles is

$$\frac{H}{\pi f}\frac{d^2\delta'}{dt^2} + \frac{E'^2}{x_{d'}}\delta' = 0$$

Calculations for a particular generator showing the phase angle of the circulating current and the natural frequency are summarized in Figs. 8 and 9 respectively. These curves show the relative effect of the circulating current in the two load division channels, namely, real and reactive current sensing, as a function of the kilowatt and kilovar system load. It is clear that the signal introduced into the load division servos by hunting of the two machines varies considerably with steady-state operating conditions.

The author mentions that the frequency of voltage modulation introduced by load transfer oscillations is just twice the frequency of load transfer. Fig. 10 gives an oscillogram illustrating this condition, which shows that the load transfer oscillation begins to build up immediately upon application of system load, and after a time a prominent second harmonic load transfer appears superposed on the original. Also the bus voltage modulation does not appear until the second harmonic load transfer has become significant. The two generator current traces show that the second harmonic modulated components reinforce one another as to current delivered to the load, while the fundamental modulation com-

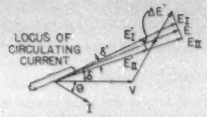


Fig. 7. Vector diagram of two equal synchronous generators in parallel with equal real and reactive loads, showing superposed hunting

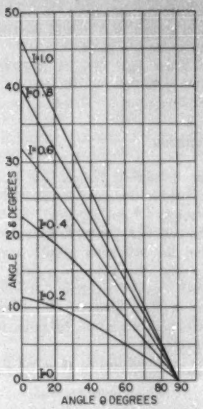


Fig. 8. Phase angle of circulating current due to hunting (equal to steady-state power angle)

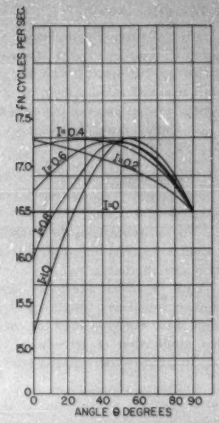


Fig. 9. Calculated natural frequency of hunting versus steady-state load per generator, 2-machine system

ponents are of a load-transfer nature. The pencil sketch envelopes at the right in the oscillogram are drawn approximately through the center lines of the second harmonic components and these envelopes emphasize the fundamental load transfer oscillation.

Therefore it is apparent that voltage modulation is due to the appearance in the load of cyclic currents contributed equally by the two machines, and these currents in turn arise as a prominent second harmonic of a fundamental load transfer oscillation. Other oscillograms show that when the load transfer oscillation is removed, the bus voltage modulation is insignificant. A study is in progress to accumulate more information of this kind and to determine analytically the reasons for this behavior.

The author has suggested certain definitions of terms and certain proposed specifications limiting the amount of load transfer and frequency modulation for aircraft a-c systems. We believe that any attempt to set these specifications generally at this time is premature and that in any case such action should be initiated through the cognizant committee of the Institute. It is suggested that any determination on specifications should be worked out together by the user, the drive manufacturer, and the generator manufacturer.

W. P. Koenig (Boeing Airplane Company, Seattle, Wash.): The problem under discussion is attaining increased importance as more and more electronic systems are being added to modern aircraft. The following is intended to expand and clarify the section of this paper which treats instantaneous load division error.

As stated in the paper, the effect of instantaneous load division error on alternator heating is so small that it may be neglected. The voltage modulation, however, which accompanies load oscillation seriously affects the performance of many of the airplane electronic systems. This voltage modulation depends upon the amplitude of the instantaneous load deviation from the average load of the individual generation unit and the rate at which the deviation occurs. If the rate is sufficiently low, it is possible that the voltage regulators will effectively compensate for the tendency for the system voltage to be modulated. If the rate is high the voltage regulators cannot compensate effectively.

The most probable frequency for load-division instability is the natural frequency of the paralleled generation units, i.e., frequency determined by the drive inertias connected by the machine reactances. In

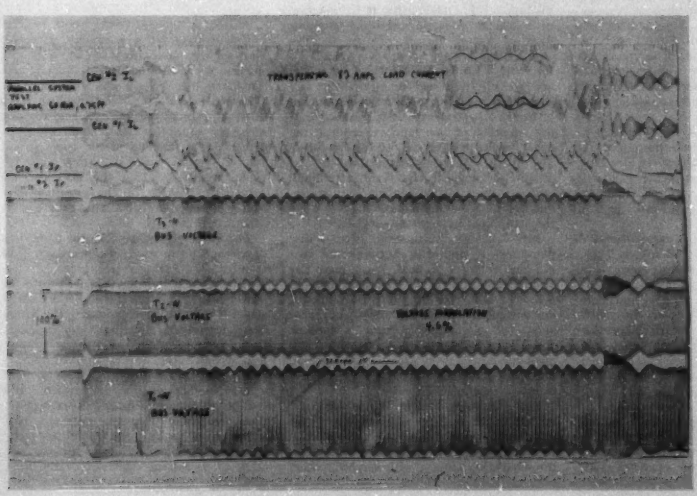


Fig. 10. Buildup of load transfer oscillation and second harmonic load modulation

practical aircraft equipment this frequency is such that the generation unit excitation system may reinforce load oscillation at this frequency.

As stated in the paper, the frequency of the voltage modulation that results from load oscillation is double the frequency of the load oscillation. For this reason, voltage modulation that results from load instability that is reinforced by the generation unit excitation system cannot be controlled by the voltage regulators.

Sustained load oscillations at frequencies higher than the natural frequency of the generation units is improbable, and lower frequencies are less serious.

Based upon experience with a large airplane, a suggested limit of load-division instability that is compatible with the other values given in this paper might be as follows:

The instantaneous load-division error shall not deviate from the average load of an alternator by more than 10% of the kilowatt rating of one alternator.

S. C. Caldwell and **R. T. Smith** (General Electric Company, Schenectady, N. Y.): Mr. Oman is to be commended for bringing this aspect of system operation to notice. We agree with him that limitations are necessary on both frequency and load division deviations.

In discussing frequency oscillations, it is of interest to note that a system can be stable as far as control response is concerned and yet appear unstable in operation due to a high noise level. If such a system has a response which settles out with three, four, or more overshoots, noise in which the disturbances are spaced on the average as close together as the response time of the controlled system will keep the system excited and ringing. The output may have components of sinusoidal oscillation at the response frequency of the controlled system. The lower trace of Fig. 2 is the sort of operation that might be expected in such a case. Another possibility is that from time to time the repetition rate of the noise disturbances may, for a short interval, coincide with the response frequency of the system. In this case, even though noise disturbances are small, sporadic oscillations may build up to high levels and die out. The lower trace of Fig. 3 is the sort of operation one might expect.

As far as the user is concerned, it makes little difference whether the source of difficulty is noise or an unstable control system if the swings in output are of a magnitude and frequency which cause trouble in the loads. However, improvement in the two situations of fluctuating output may take quite different paths. In a noisy but stable system, additional damping may be introduced to reduce output fluctuations at the expense, usually, of response speed and increases in complexity, size, and cost. Clearly, the more direct, if not more difficult, approach is to reduce the noise generated in the system. In the case of an unstable control system, of course, the method of increased damping must be used.

In parallel operation of systems in which noise is a sizable factor, fluctuations take three forms:

1. Load oscillations may take place at any of the quasi-resonant frequencies of the parallel system. Often, two quasi-resonant frequencies may exist, one at a frequency determined largely by generator and regulator characteristics together with system inertias, and another one which is slower, and in which generator and regulator are essentially in the steady-state region, and the speed-governing system together with real load division characteristics are the ruling factors. This is pronounced in hydraulic drives in general use but will not necessarily be true when faster governing and load division systems are used.

2. Bus frequency (neglecting impedances between two machine busses) will swing with the noise disturbances but will tend to be smaller in amplitude than if the machines are operating isolated.

3. As pointed out by the author, bus voltage will also swing both in the unstable case and in the noise case. The frequency of voltage fluctuation may be either at the same frequency as load swings or may be at double frequency, depending upon the rate of load variation, reactive division sensitivity to real load, regulator load regulation, and speed of response of the regulator compared to load variation rate.

The introduction of double-frequency voltage fluctuations is an interesting case. The author helpfully provided details of the test results and the arrangement of the equipment. Without going into an extensive analysis, we feel that the double frequency may result from the following. Consider the diagram in Fig. 11.

For $|E_1| = |E_2|$ it is seen that, as E_1 and E_2 move apart as shown, V_T will decrease from a value equal to $|E_1|$ to $|E_1| \cos \delta$. Then, as E_1 and E_2 swing back together, V_T increases to a value equal to $|E_1|$ again. When E_1 and E_2 swing past each other on the second half of the cycle of rotor oscillation, V_T again decreases to $|E_1| \cos \delta$. Thus V_T swings twice for each cycle of rotor angle. Mathematically, if δ is a small angle which varies sinusoidally

$$\delta = K \sin \alpha t \tag{1}$$

where

$\alpha = 2\pi$ times the frequency of rotor oscillation

and

$$\cos \delta \approx 1 - \frac{\delta^2}{2} \tag{2}$$

When $|E_1| \cos \delta$ is expressed in terms of equations 1 and 2

$$V_T = |E_1| \cos \delta \approx |E_1| \left(1 - \frac{K^2}{4} + \frac{K^2}{4} \cos 2\alpha t \right)$$

and it is seen that V_T does vary at twice the frequency of δ (which determines the frequency of load oscillation).

At load oscillation frequencies of under about one cycle per second, the regulator will follow the voltage fluctuations and reduce the double-frequency component to small values. However, if impedance exists between the regulating points, and reactive division circuitry is sensitive to real power, voltage modulation at fundamental load oscillation frequency will occur. At higher frequencies of load oscillation, excitation

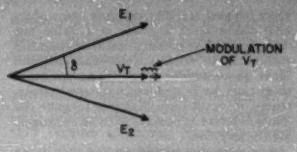


Fig. 11. Voltage modulation

system ability to follow and reduce bus voltage fluctuations decreases, and the double-frequency component becomes larger up to the point where no changes in excitation are provided by the excitation system.

It is of interest to note that in the transitional frequencies between those in which the excitation system follows voltage fluctuations and those in which it cannot follow, the double-frequency voltage variation may actually be increased over that obtained at the higher frequencies if the excitation system is lightly damped.

Noise which is not offensive in isolated operation may easily cause objectionably large load oscillations in parallel systems. Since power transfer is based upon machine impedance approaching transient reactance, a very small swing of the rotor on one machine will cause a large transfer of power. As an example, let

$X_d' = 0.3$ pu (per unit)

and rpm = 6,000

Then, a difference in mechanical angle of about three degrees will produce a power transfer of about 0.3 pu. If the rotor is oscillating at a rate of five cycles per second with respect to the rotating axis average speed, this represents an amplitude of speed variation of roughly 0.0025 pu, which corresponds to but one cycle variation in frequency in isolated operation. In practical systems, this difference is not as noticeable as in this example, since the transfer of power serves to reduce the amplitude of rotor swing; hence, the numbers used by the author in defining frequency deviation limits of one cycle for isolated operation may be reasonably consistent with peak power interchanges of 0.2 pu for the parallel case when the five-cycle rate is encountered. When other load oscillation frequencies are considered, these numbers may not be consistent.

The author in presenting his sample specification opens an area which will require a great deal of discussion before a more or less standard specification can be derived to permit the equipment manufacturer to use the same design on most airplanes. Our feeling is that permissible amplitude versus frequency plots for speed, load division, and voltage modulation would be extremely useful to equipment manufacturers and would be an adjunct to permissible limits of instantaneous and average deviations.

We do not believe it is reasonable to use a linear relationship between deviation limits under rated conditions and overload conditions. In parallel operation, many systems

show reduced stability margins under overload conditions: they are still stable but may deviate more than a linear relationship with normal conditions would indicate. Since these overload conditions are on a short-time basis and are not usual occurrences, it would seem desirable to make the limits as broad as possible without endangering the airplane.

As a general comment, we feel that limits should be set on the operation of this equipment. This is desirable to obtain better performance on sensitive load equipments in normal operation. However, these limits do not ensure that even worse fluctuations than those described by the author will not exist in flight as a result of malfunction of generation control or drive equipment. Furthermore, system protection against such fluctuations is, at best, incomplete. Therefore it seems that critical items of equipment such as autopilots should be able to tolerate fluctuations up to the point at which system protection will correct the trouble without endangering the airplane (by possible violent operation of airplane controls in the case of autopilots).

Erle P. Barlow (Glenn L. Martin Company, Baltimore, Md.): This useful paper draws attention to a practical problem. The author's example of a specification for frequency stability directly limits the magnitude of the instantaneous frequency variation on both a short-time and a long-time basis. The type of system covered by the example when limited to so small a frequency deviation would not be likely to cause difficulty due to any high rate of variation within the prescribed tolerance band. Other means of generating 400-cycle power, such as a transistorized oscillator, will not be limited by generator rotor and drive inertia; hence, frequency change rates considerably higher than those considered by the author will be possible. Therefore a general definition of frequency modulation should include the effect of rate of frequency change. Further, the number of occurrences of such disturbances per unit time is significant; for example, a transient which disturbs the autopilot at some noncritical time, once per flight, is probably acceptable while a more frequent disturbance occurring at an unpredictable time, perhaps during a bomb release run, would probably be unacceptable. It is therefore suggested that a general definition of frequency modulation should include magnitude of frequency variation, rate of frequency variation, and number of occurrences of the variation per unit time. For a periodic type of modulation, the modulation frequency may be a more convenient expression than rate of frequency variation.

Henry Oman: The discussions all contribute to a better understanding of frequency and load-division instabilities and are much appreciated. It is encouraging to note from Mr. Barlow's discussion that the Air Transportation Committee is to formulate definitions for frequency modulation.

ADDITIONAL DISCUSSION

The following discussion was received after the paper had been re-presented at the 1956 AIEE Winter General Meeting.

Harmonics from Railroad Rectifiers on Power System Reduced by Filters

Discussion of paper 55-627 by S. J. Bozzella, J. L. Kennedy, M. Mahr, Jr., and H. W. Wahlquist, published in Applications and Industry, November 1955, pages 324-34.

E. B. King (American Telephone and Telegraph Company, New York, N. Y.): Although filters and resonant shunts have previously been installed at a number of rectifier locations in the United States, several unique problems were encountered and overcome in the installations discussed by the authors to make this a most interesting paper which will be a valuable addition to the available information. The authors are to be congratulated on the manner in which a difficult problem was handled, as well as on the excellent discussion in the paper.

Some examples of the unusual features of this project and the unique methods of handling the problems are summarized as follows:

1. It was impracticable to use multiphase operation of the rectifiers to reduce the power system influence, due to the use of rectifier equipment that had previously been used for another purpose in a war plant.
2. The rectifier supply circuit was also used as an extensive customer distribution circuit, which was involved in considerable joint use with New York Telephone Company communication circuits.
3. Rapidly changing and exceptionally high voltage telephone influence factors (TIF) in the range of 100 to more than 300 were present on the 13-kv distribution circuits before remedial measures were applied.
4. Standard power factor correction capacitors and reactors were used in the filters instead of precision types previously used in most rectifier problems.
5. To my knowledge this is the first installation of standard distribution transformers with ratings as high as 15 and 25 kva to be used as reactors in connection with filters for reducing power system influence. A clever method was employed to reduce the heating effects caused by hot spots which developed in the transformer case opposite the air gap during heavy loads in the rectifier.

W. A. Munson and **R. W. Ferguson** (Westinghouse Electric Corporation, East Pittsburgh, Pa.): This paper was reviewed with considerable interest as it included data on the system both before and after applying harmonic filters as well as all pertinent information on the filters themselves. Thus, for comparison purposes, an analysis was made using the theoretical values of harmonic currents as determined from the "Electrical Transmission and Distribution Reference Book."[1]

To enable using the information published in the reference, the following assumptions were made:

$IX/E = 0.08$ and $GCR = 1.0$

It was further assumed that the harmonic impedance of the 66-kv system, shown in Fig. 1 of the paper, was 0; i.e., for harmonic currents, the 66-kv system was the equivalent of a ground point.

Using the foregoing assumptions and harmonic currents obtained by use of the reference, the voltage TIF was calculated for the condition of Babylon Yards Substation only in operation without any filter. This calculated voltage TIF was found to be 186.2 whereas Table V of the paper gave a root-sum-square (RSS) value of 100. The filter was then assumed connected and the voltage TIF at the filter terminals was again calculated. Here, the calculated value was 33.9 versus a reported value of 25.

The process of comparison was continued for the case of the Lindenhurst Substation only in operation. For this condition the calculated voltage TIF's were as follows:

Without any filter in operation: Voltage TIF was 128

With 800-cps filter only: Voltage TIF was 148

With 420 800-cps filters: Voltage TIF was 41.9

With 300 420- and 800-cps filters: Voltage TIF was 33.3

These values of voltage TIF compare as follows with the reported values in Table VI of the paper:

Theoretical Calc.	RSS	Measured
128	versus 120	
148	versus 63	63
41.9	versus 40	38
33.3	versus 20	21

These comparisons between the theoretical calculated values of voltage TIF and the actual measured results indicate that if harmonic filters are applied on the basis of the theoretical values, the resultant TIF value of the system will be less than calculated.

Also, these comparisons indicate that the theoretical values of harmonic content are pessimistic and can be safely used for filter application purposes. It must be remembered, though, that the effectiveness of any calculated filter is only going to be as good as the system information used for selection of filter size.

Reference

1. Electrical Transmission and Distribution Reference Book (book). Westinghouse Electric Corporation, Pittsburgh, Pa., fourth edition, 1950, chap. 23.

High-Potential Testing of Aircraft Equipment Electric Insulation

L. B. KILMAN
MEMBER AIEE

J. P. DALLAS
MEMBER AIEE

Synopsis: The purpose of equipment dielectric test is to detect insulation faults, not to stress solid insulation. An investigation of air-gap breakdown through holes in aircraft insulation materials simulating faults in aircraft equipment demonstrated the following: 1. the 500-volt high-potential tests will not detect common faults in aircraft insulation; 2. the common concern over insulation damage from repeated high-potential tests is not justified; 3. dielectric breakdown through small diameter faults in insulation is more likely to produce immediate failure than similar discharges over surface paths. A single circuit interruption transient discharging through a 0.003-inch-diameter hole may completely fault the insulation, while a discharge of similar energy across surface gaps may require hundreds of repetitions before failure; and 4. when small-hole discharges through insulation occur between copper and iron, insulation may fail immediately even with supposedly tracking resistant materials such as mica. Insulation puncturing and tracking circuit interruption high-voltage transients reaching values of several thousand volts are discussed. The value of these transients is shown to be substantially the same for 28-volt systems as for higher voltage systems and the critical circuit factors determining the value of these transients are analyzed.

THERE are two ways of finding out whether aircraft equipment insulation is adequate: one is to wait for service failures and the other is to test it properly. The high-potential dielectric test is the principal means of detecting holes, cracks, discontinuities, metal inclusions, and other flaws in insulation. Because all insulating materials contain flaws, are subject to damage in fabrication, deteriorate with age, and because material inspection is never 100 per cent perfect, properly engineered electric insulation can and does fail.

The detection of an insulation fault at the earliest possible stage minimizes the cost and hazard. An inspection-defying 0.003-inch-diameter pin hole, or crack where the insulation was intended to lap, if detected when the coil is first assembled on the core will cost the least to correct. If the test is not made at this point, additional man-hours invested in this faulty part will cause the total cost to go up and up. After this faulty insulated coil is impregnated and baked, after assembly of the complete machine, at the customers receiving inspection, at each stage the cost is multiplied. Finally if this insulation fault is not discovered until failure in flight, truly high penalties may result.

Rigorous insulation quality control at any or all preceding stages may not eliminate the need for high-potential testing at later stages. In the progress of an aircraft electric component from the basic materials through usage to the scrap heap, the various stages of manufacture and usage all introduce the possibility of insulation damage or deterioration. The only sure way to determine if the insulation is airworthy is to test it.

Proper High-Potential Tests Are not Injurious to Acceptable Insulation

The common concern over insulation damage is not justified when proper high-potential test methods are used on equipment insulation which is up to acceptable standards. Normal high-potential tests of equipment are never intended to determine the maximum voltage which the insulation will withstand. Dielectric strength tests are normally applied to insulation materials but not to fabricated equipment. The entire purpose of equipment dielectric tests should be to discover flaws in the insulation. Therefore, a high-potential test voltage should be high enough to discharge through a hole in the insulation but not high enough to stress seriously the solid insulation materials or surface air gaps designed to acceptable aircraft dimensions. The better aircraft insulation materials have breakdown potentials in excess of 1,000 volts per mil. It is obvious if 0.003-inch insulating material contains a pin hole that the potential required to break down this air gap will not seriously stress either the solid insulation or normal air gaps in the equipment. Marginally thin insulation will however be detected. A voltage sufficient to breakdown a 0.003-inch air gap may puncture insulation which is dangerously thin. Solid insulation can be dangerously thin because switching transients often cause high voltages in the normal operation of the equipment.

Because inductive circuit interruption transient voltages may reach values of more than 1,000 volts, it is indeed important that high-potential test voltages should detect holes, cracks, and marginally thin insulation. If these discontinuities are not detected by high-potential test, then, in operation, inductive circuit interruption transients are very likely to discharge through these flaws or puncture marginally thin insulation and cause insulation failure by tracking.

Dielectric Fatigue

Repeated dielectric stresses do produce voltage fatigue failure of insulation. However, the conclusion that repeated high-potential tests of equipment are therefore damaging and undesirable is not justified. In the first place, this phenomenon is analogous to fatigue stress in metals. There is a voltage fatigue limit effect, with voltages below this limit in value producing a negligible effect on insulation. While far too little is known about voltage fatigue limit values for aircraft insulation, it is safe to say that proper dielectric test procedure will not even closely approach this limit for acceptable thickness of insulation where the test voltage has been selected to discover holes in the insulation and, therefore is maintained substantially below the ultimate dielectric strength of the material.

An illustration may be in order. The dielectric strength of a 3-mil-thick section of Mylar at 150 degrees centigrade temperature is given as 3,000 volts per mil. Dielectric fatigue effects for 10,000 or less repetitive voltage stresses disappeared at approximately 80 per cent of the breakdown voltage for this material, in this case, 9,000 volts. Therefore, repeated dielectric tests at less than 7,000 volts would be unlikely to affect a measurable reduction in the breakdown voltage of this insulation. When it is considered that the proper dielectric test for this insulation would be 2,000 volts, or less, and similar stresses to this might occur in operation as a result of switching transients, it seems unjustified to limit the use of the high-potential test technique on the basis that insulation damage is likely to occur.[1-4]

Paper 56-441, recommended by the AIEE Air Transportation Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE South West District Meeting, Dallas, Tex., April 3-4, 1956. Manuscript submitted January 11, 1956; made available for printing February 16, 1956.

L. B. Kilman and J. P. Dallas are with the Hughes Tool Company, Culver City, Calif.

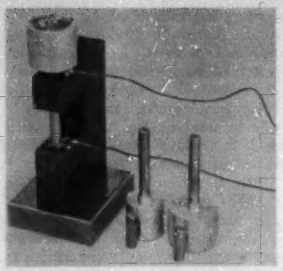


Fig. 1. Test jig used for testing breakdown voltage and tracking of insulation

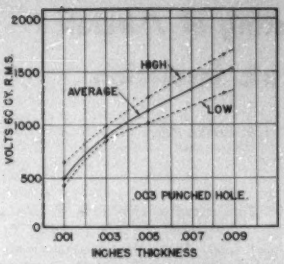


Fig. 2. Dielectric breakdown voltage of holes in 0.003-inch-thick Mylar insulation. Ten high-potential tests made at each indicated thickness

Solid line—average breakdown voltage
Dotted lines—maximum and minimum breakdown voltage

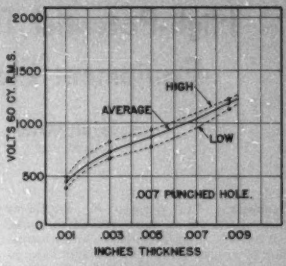


Fig. 3. Dielectric breakdown voltage of holes in 0.007-inch-thick Mylar insulation. Ten high-potential tests made at each indicated thickness

Solid line—average breakdown voltage
Dotted lines—maximum and minimum breakdown voltage

Minimum Test Voltage to Detect Flaws in Equipment Insulation

The 500-Volt High-Potential Test Was Inadequate

Failure analysis of 28-volt equipment repeatedly demonstrated that 500-volt 60-cycle a-c tests were not detecting insulation flaws which subsequently cause service failures. Circuit interruption transient voltages discharged across the insulation discontinuities which the standard 500-volt test potential had failed to discover. It was realized that simple calculations made on the basis of the air gap demonstrated that 500-volt high-potential tests were not enough. Existing information could have provided the basis for approximate determinations of dielectric test voltages, but these left many unanswered questions.

Breakdown Voltage Through Insulation Holes

Because investigation of available literature on the subject did not uncover usable information concerning breakdown voltages of air gaps through small-diameter openings in present aircraft insulation, an experimental investigation of this subject was undertaken. It was anticipated that the following factors might alter or affect the breakdown potential through insulation flaws:

1. Voltage gradients in the gap area caused by the change in dielectric constant from insulating material to air at the edges of the hole.

2. Limited dissipation of the heat of discharge caused by loss of convection cooling and surrounding the gap with insulating material.

3. Contamination of the air-gap area with products of the discharge from the metals and insulating materials.

4. Air-gap pressure changes caused by confining the discharge in small holes.

While individual analysis of these and other factors possibly affecting air-gap breakdown through small holes in insulation was beyond the scope of this investigation, an over-all experimental determination of the combined effects was made. Various sizes of holes in a range of commonly used thicknesses of Mylar, Teflon, mica, and vulcanized fibre (fish paper) were tested for breakdown voltage under carefully controlled conditions. Hole diameters used were of 0.003, 0.005, 0.007, and 0.009 inch. Insulation material thickness was varied for each size of hole from 0.001 to 0.009 inch in commercially obtainable steps. The test jig shown in Fig. 1 was constructed in accordance with American Society for Testing Materials (ASTM) *D149-44* specifications[5] for the test of insulating tapes with opposing cylindrical rod electrodes 1/4 inch in diameter with edges rounded to a radius of 1/32 inch resting on the insulation with a force of 1/10 pound. The electrode surfaces were polished to a 10-microinch finish and were repolished after each discharge. It was found that consistent results could not be obtained unless extreme care was used to avoid contamination of electrode and insulation surfaces. A fingerprint could alter results beyond use. Medical operating theater type of care and cleanliness were required. Since the dielectric breakdown voltage of the air gap in the holes in the insulation was under investigation temperature and humidity changes had no significant affect on the result. To check this point, similar voltage breakdown tests of 0.003-inch holes in 0.003-inch Mylar tape were made at 125 degrees centigrade and at room temperature. The differences between the two series of tests were all within the ±20-per-cent repeatability range of the test procedure. To those not familiar with this type of test, a ±20-per-cent repeatability of air-gap breakdown voltage data is not easy to achieve. The technique of punching uniform diameter holes without thickening the insulation around the hole in materials which tend to flow, such as Mylar, is a real problem complicated beyond the scope of this discussion. It should be noted, however, that microscopic measurement of each hole diameter and thickness must be made to secure repeatable results.

Dielectric breakdown voltage of holes in Mylar insulation, shown in Figs. 2 and 3, is typical of all the insulating materials. It is to be noted that a 500-volt rms 60-cycle high-potential test currently used on much aircraft equipment would not detect a flaw reliably in insulation thicker than 0.001 inch. One thousand rms volts would be required to detect a 0.003-inch-diameter hole in 0.003 inch of insulation. As had been expected, larger diameter holes broke down at lower voltages.

Insulation Failure Caused by Circuit Interruption Voltage Transient Discharge Through Insulation Holes

An Alarming Observation

To relate the observed facts directly to aircraft equipment operation a typical aircraft circuit was set up so that the circuit interruption transient voltage could be applied directly to a 0.003-inch hole in 0.002-inch-thick Mylar insulation under the conditions specified in the fore-

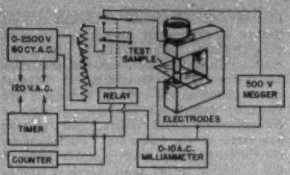


Fig. 4. Schematic diagram of test setup for testing breakdown voltage and tracking of insulation

going. A standard 28-volt fire-extinguisher solenoid producing a maximum circuit interruption 1,000-volt transient was applied to samples in the jig shown in Fig. 1. Complete tracking to a low resistance below 1,000 ohms occurred in a single discharge. When this test was repeated with 20 fresh samples, in no case did it take more than 10 discharges to produce insulation failure.

The seriousness of this observation is indicated by the fact that it took 100 to 200 similar discharges across the machined surface of a fabric-base aircraft-quality phenolic insulation to produce a similar failure. The energy in the circuit interruption voltage transient of this fire-extinguisher solenoid was 7 watt-seconds, but the maximum rate of energy dissipation was 338 watts. It is the high rate of energy dissipation in circuit interruption transient spark-overs that is so damaging to insulation. The electrodes used during this part of the test were made of brass. It appeared that either Mylar had exceptionally bad tracking characteristics or that the tendency to produce tracking was greatly increased when the discharge was confined to a small, enclosed area.

Investigation of small-hole tracking of some insulation materials was undertaken to answer the questions raised in connection with the fire-extinguisher solenoid switching transient test described in the foregoing. Referring to the schematic circuit diagram of Fig. 4, the test jig of Fig. 1 was connected to a standard 2,500-volt high-potential dielectric test instrument which provided close control of both voltage and current. Breakdown voltage was continuously controllable from zero to 2,500 volts, and the discharge current was continuously controllable from 50 microamperes to 20 milliamperes. An automatic timer was connected so that discharge time was controlled for $2^1/_2$-second ON and $2^1/_2$-second OFF periods. This timing cycle allowed stable meter readings to be taken before, during, and after each discharge period. The deviation from ASTM *D495-42* arc resistance test method was made to adapt the procedure to small-hole discharge conditions. The ASTM specified 10-milliampere current proved to be in excess of values which would allow distinguishing between the small-hole tracking characteristics of the various insulation materials. In this test it was chosen to vary the current below 10 milliamperes rather than the time of discharge. A control relay connected a 500-volt standard megger to the dielectric immediately after the high voltage had been cut off. Tungsten electrodes are suggested for arc resistance tests in the ASTM *D495-42* specification, so the test jig in Fig. 1 was initially set up with tungsten electrodes of the same dimensions as before; i.e., with a 1/4-inch diameter with 1/32-inch radius on the material-contacting ends and a 10-micro-inch polish. Because results with the tungsten electrodes did not correlate with the brass electrodes first used, the subject of electrode material effect was checked. Electrodes of copper and iron were also used to simulate more closely the actual breakdown conditions in aircraft equipment.

Electrode Materials Critical

Referring to Table I, it will be noted that a 4-milliampere discharge through a 0.003-inch hole in Teflon could be repeated for 500 $2^1/_2$-second periods with tungsten electrodes, while a similar discharge using iron and copper electrodes failrd the insulation in four $2^1/_2$-second discharges. Mylar and mica also showed a drastic reduction in small-hole tracking resistance when copper and iron electrodes were used. Since the tracking resistance of mica is generally considered to be very high, it was surprising to find a 4-milliampere discharge failing the insulation in ten $2^1/_2$-second cycles. Under the same conditions with tungsten electrodes, the mica holes were able to withstand 200 $2^1/_2$-second discharge cycles. From the data obtained in this test the following was learned:

1. The electrode metals for small-hole tracking tests are critical and, if the results are to correlate with operating experience, the electrode materials must more closely correspond to those in the equipment.

2. Tracking failure through flaws in insulation occurs much more readily than over the surface of insulations.

Table I. Insulation Failure Through 0.003-Inch-Diameter Holes Showing Effect of Electrode Materials

Insulation Material	Current in Breakdown, Milliamperes	Number of Cycles before Insulation Failure		
		Tungsten–Tungsten Electrode	Iron–Copper Electrode	Iron–Iron Electrode
Mylar	0.031		31	
	0.031		1	
	0.031		9	
	0.062		3	27
	0.062		4	11
	0.062		1	19
	0.062		2	
	0.125			8
	0.125			5
	0.25			3
	0.5	150*		1
	1.0			9
	1.0			1
	1.5			2
	2.0	250*	1	
	2.0		1	
	2.0		1	
	3.0	350*		
	3.0	500		
Teflon	0.5		90	
	0.5		202	
	0.5		250	
	1.0		90	
	1.0		60	
	2.0		15	
	4.0	500*	4	
	4.0	500*		
Mica	0.5		28	
	0.5		28	
	1.0	200*	25	
	1.0	200*	20	
	2.0	500*	21	
	2.0	215*	16	
	4.0	187	10	
	4.0	208	11	
	20.0	8	2	
	20.0	14	4	

Note: All holes = 0.003-inch diameter.
One cycle = $2^1/_2$ seconds ON and $2^1/_2$ seconds OFF.
* Did not fail.

3. Copper and iron electrode surfaces in a discharge through small holes in insulation will cause tracking resistant materials such as mica to fail rapidly.

4. Common circuit interruption switching transients contain enough energy to fail insulation immediately when discharging through small-diameter holes.

Important Background Factors

The effect on aircraft equipment reliability of this small-hole insulation breakdown and tracking phenomenon is best appraised against a background review of the factors causing electric insulation deterioration and circuit interruption transient voltage effects.

Deterioration of Electric Insulation

The fact that an insulation, at some stages of its manufacture, has passed adequate inspection test gives no assurance that even in storage the insulation will not have deteriorated to the point where the equipment is no longer airworthy. Some of the factors causing electric insulation deterioration are the following:

1. Time-electrochemical reactions with air, etc.

2. Evaporation of plasticizers from plastic insulation materials which cause embrittlement leading to subsequent vibration failure of the insulation.

3. Moisture is more likely to damage nonoperative equipment since the heat of operation tends to dry out operative electric equipment.

4. Electrolytic corrosive action to which metals are subject tends also to promote the deterioration of electric insulation in the vicinity where the corrosive actions occur because of the metallic salts generated. Many insulating materials themselves, when subjected to potential differences in the presence of high humidity, rapidly deteriorate.

5. High temperature deteriorates electric insulation. Heat greatly accelerates the changes mentioned in the foregoing.

6. Thermoplasticity and cold-flow effects due to temperature and pressure often cause thinning or flowing of insulation which, to an extent, produce ultimate failure of an originally acceptable insulation thickness. Thermoplastic and cold-flow effects are a function of time as well as a function of temperature and may occur very slowly in operation.

7. Inductive circuit interruption voltage transients may reach values of more than 1,000 volts. These voltage surges, when present, are likely to puncture insulation if winding pressures, thermoplasticity, or cold-flow effects have reduced insulation thickness below critical values. For many commonly used insulation materials and circuits, insulation thickness of less than 0.002 inch may be critical. Failure in 28-volt circuits is not immediate from such punctures, but ordinarily occurs as the result of repeated interruption transient discharge through the initial puncture causing tracking of the insulation around the edges of the hole. Proper high-potential test procedure will often discover these usage-developed faults before there is any failure.

8. Tracking, or the build-up of a conductive path across surface gaps caused by carbonization of an insulator by repeated circuit interruption inductive surge discharges, is a major cause of insulation failures.

9. Corona discharges rapidly deteriorate some types of commonly used aircraft electric insulation. At high altitude, corona discharges occur at much lower voltages than at sea level, which makes this consideration an important factor in aircraft insulation deterioration.

10. Surface contamination.

Circuit interruption voltage transients severely stress the insulation of not only the inductive items of equipment generating the surge but also the cable, terminals, connectors, resistors, capacitors, etc., in the circuit from the hot side of the inductance to the load side of the switch opening the circuit. Thus, noninductive items of equipment used in inductive circuits may be affected.

Phillips and Mitchel[2] presented substantial data showing transient voltages in excess of 3,000 volts for 28-volt circuits interrupted with a snap-action double-break air-gap switch. A detail investigation of this problem has resulted in the following observations:

1. A typical interruption cycle for an aircraft circuit may be of 10-millisecond duration. An arc is initially formed as the contacts open. During the period of this arc, the current falls at a relatively slow rate until a critical arc current value is reached below which a gaseous discharge will not be maintained. At this point, an abrupt discontinuity in the current flowing in the circuit will occur as the discharge extinguishes. It is this current discontinuity and the accompanying rapid change in flux which produce the interruption voltage transients under consideration. If the total circuit interruption time is 10 milliseconds, it is likely that this final voltage transient producing current flow discontinuity will be a few microseconds. The resulting high-voltage microsecond transient may ionize the gas in the switch gap and restrike the arc, repeating the foregoing cycle.

2. When the circuit interruption transient is computed from the familiar relation $e = -N\ d\phi/dt\ 10^{-8}$ volts or any other derivation of Faraday's law, an improper conclusion may result. Computation on the basis of any total circuit interruption time of the order of milliseconds will produce a misleading low figure likely to be as small as 10 per cent of the probable actual circuit peak values. The possible peak transient values may be computed for the microsecond periods of abrupt current change in the circuit interruption current characteristic. When this is done, an equally misleading high value of the order of 5,000 to 10,000 volts is obtained. Actual circuit interruption voltage transients are normally neither as low nor as high as computation would indicate for reasons suggested in the following. Furthermore, computation of the electromotive force of self-induction rather than practical laboratory investigation has led to the false conclusion that switching voltage transients will vary as the system voltage. This spurious conclusion has been used to support the establishment of various high-potential test voltages for comparatively small differences in nominal system voltage such, as a 500-volt high-potential test for 28-volt systems.

3. Actual circuit interruption voltage transient values are determined by the minimum instantaneous voltage breakdown at the weakest point in the circuit from the inductive source of energy to the switch gap. In a properly designed circuit, the breakdown voltage of the switch gap at the instant of circuit interruption discontinuity should be less than the breakdown voltage of any other voltage transient stressed part of the circuit.

4. Therefore, 28-volt system voltage transients will not differ substantially from other high-voltage aircraft systems since in every case the actual observed values are really determined by voltage breakdown of the switch gap or some other part of the circuit. As pointed out in the foregoing, the theoretical values computed from the maximum rate of current change are likely to be above 5,000 volts so that it is always the switch gap or other insulation breakdown that determines the actual circuit values measured. While the static voltage breakdown values of switch gear for the various systems may be substantially different, the operating condition at or near deionization of the switch gaps is not likely to differ substantially.

5. Measurement of the circuit interruption transient voltage requires the use of oscilloscopes, voltage dividers, and recording cameras capable of calibrated response to higher than 1 megacycle.

6. High frequency as well as high voltage of circuit interruption voltage transients is a key factor in measurement control and understanding of these phenomena. With voltage changes of the order of 100 volts per microsecond, it is obvious that circuit capacitance, dielectric losses, and core losses are key factors.

References

1. Report of Advisory Staff for Aircraft Electrical Systems, and Appendices. WADC and BuAer, Wright Air Development Center, Dayton, Ohio, U. S. Navy Bureau of Aeronautics, Washington, D. C., Apr. 1951.

2. Peak Voltages With D-C Arc Interruption for Aircraft, Virgel E. Phillips, Walter P. Mitchel. *AIEE Transactions*, vol. 63, Dec. 1944, pp. 944–49.

3. Test Code for D-C Aircraft Rotating Machines. *AIEE Test Code No. 800*, Jan. 15, 1953.

4. Comments on Aircraft Switch Testing T. R. Stuelpnagel, J. P. Dallas. *AIEE Trans.*

actions (*Electrical Engineering*), vol. 73, pt. II, Nov. 1954, pp. 301–05.

5. Standard Methods of Test for Dielectric Strength of Electrical Insulating Materials at Commercial Power Frequencies. *Specification D149-44*, American Society for Testing Materials, Philadelphia, Pa., 1946.

6. High-Potential Dielectric-Strength Testing, John Keefe. *Electrical Manufacturing*, New York, N. Y., May 1935.

7. D-C Arc Interruption for Aircraft, J. S. Quill, L. T. Rader. *AIEE Transactions*, vol. 63, Dec. 1944, pp. 883–88.

8. Effect of Electric Discharges on the Breakdown of Solid Insulation, T. W. Dakin, H. M. Philofsky, W. C. Divens. *Ibid.*, pt. I, May 1954, pp. 155–62.

9. Electrical Insulation (book), G. L. Moses. McGraw-Hill Book Company, Inc., New York, N. Y., 1951.

Discussion

D. W. Exner (Boeing Airplane Company, Seattle, Wash.): This is, in my opinion, one of the more significant papers which has been presented in recent years on aircraft equipment problems. The experimental data presented by the authors show quite clearly that the conventional 500-volt dielectric test, so widely specified for low-voltage equipment, will not reliably detect minute flaws in the thinnest practical homogenous insulations. Unless such flaws are detected, equipment exposed in service to aircraft environments involving wide altitude and temperature cycling, with consequent condensation and penetration of moisture, will fail prematurely because of electrolytic deterioration or by direct breakdown and tracking. The problem is aggravated by the switching surge voltages described in the references of the paper.

To cite an example of typical failure to appreciate aircraft application problems, a certain instrument committee is recommending a dielectric test which calls for a test voltage in the range of "five times nominal circuit voltage." On a 30-volt device, this would call for a 150-volt dielectric test! The devices in question are important components in servo systems which are vital to the performance, and sometimes to the safety, of airplanes! It is time that insulation requirements are viewed with greater realism.

Development of Definitions for Voltage Modulation Characteristics in 380- to 420-Cycle-Per-Second Aircraft Electric Systems

O. MARKOWITZ
MEMBER AIEE

Synopsis: Details are given of the work accomplished at the Naval Air Development Center (NADEVCEN) to develop an adequate definition of voltage modulation in aircraft constant-frequency electric systems. Instrumentation was developed for the laboratory phase of the investigation. From the laboratory data and other commercial correlating data, a new and more complete definition of voltage modulation was developed. This definition is also recommended for integration into other military specifications involving aircraft a-c electric generating and regulating equipment.

Paper 56-434, recommended by the AIEE Air Transportation Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE South West District Meeting, Dallas, Tex., April 2–4, 1956. Manuscript submitted November 14, 1955; made available for printing February 14, 1956.

O. Markowitz is with the U. S. Naval Air Development Center, Johnsville, Pa.

The laboratory work in obtaining the material for this paper was accomplished by Milton Utley. The author gratefully acknowledges the information he received on voltage modulation from equipment manufacturers and aircraft companies for co-ordination with the Naval Air Development Center efforts. The author also wishes to thank the many engineering groups throughout the country who have been extremely co-operative.

The opinions or assertions contained herein are the author's and are not to be construed as official or reflecting the views of the Navy Department or the Naval establishment.

LITTLE written information is available about the characteristics of voltage modulation in aircraft a-c electric systems, and it is difficult for designers of air-borne electric or electronic utilization equipment to recognize all the effects of voltage modulation during design stages. The net results can be a lack of compatibility between utilization equipment and aircraft electric systems at their meeting point in an aircraft.

There are military specifications[1,2,3] which define the limit of voltage modulation as 2 per cent. However, the frequency distribution of the voltage modulation has not been recognized or defined. Until the characteristics of voltage modulation become more confining by specification definition, aircraft electric system designers cannot be fully aware of the effects of voltage modulation nor can any improvement be anticipated. This report attempts at such definition from data obtained in the NADEVCEN laboratories plus correlation with data obtained from other engineering groups. The scope of this paper is limited to the so-called constant-frequency systems between 380 and 420 cps (cycles per second), which are obtained from turboprop engines, direct-driven generators, and constant-speed drives, with other engines and inverters. Generators directly driven from reciprocating engines resulting in wide frequency range systems are beyond the scope of this paper.

Instrumentation

Voltage modulation is only a small percentage change in the output voltage; therefore, it is necessary to use a method that will measure small differences of much larger values. Measuring the level of voltage directly at its highest and lowest levels and subtracting for the difference to determine the extent of voltage modulation would introduce large errors. It has been the general practice to use a-c half-wave voltage, then buck out all but the variation in voltage and, with an accurate direct-voltage change, measure the variation directly in terms of peak-to-valley voltage E_{p-v} of the variation alone. If the d-c bucking voltage E_s were coincident with the minimum point of the voltage modulation, then the percentage of voltage modulation may be expressed as

$$\%\text{MOD} = \frac{E_{p-v}}{2E_s + E_{p-v}} \, 100$$

Instrumentation was set up with a variable d-c bucking voltage adjustable to determine the points of coincidence with the lowest and highest peak values of the voltage modulation. It may also be set up with a fixed d-c bucking voltage close to but slightly under the minimum peak and the remainder recorded or displayed on a calibrated scale. The method used to set up the instrumentation at NADEVCEN is shown in Fig. 1. This setup was very flexible in that the

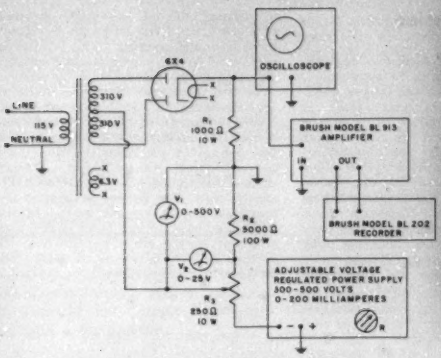


Fig. 1. Instrumentation to measure amplitude and frequency characteristics of voltage modulation

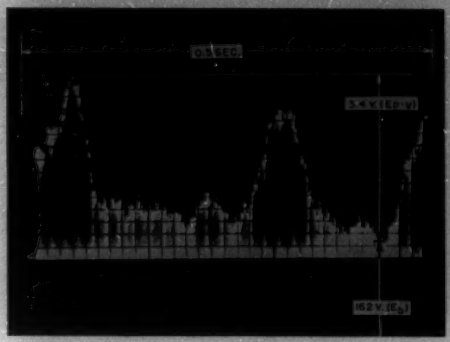


Fig. 2. Oscillograph showing complex frequency characteristics of a voltage modulation envelope

minimum bias voltage E_B could be set by R in the regulated power supply while R_3 was at zero, and measured directly on V_1. Then, the incremental change to increase the bias voltage was set with R_3, and V_2 gave a direct reading of the modulation component E_{P-V} when a factor caused by the transformer ratio was introduced. The indication for proper setting of the bias voltages was determined by observing the oscilloscope for coincidence of the bias at the minimum modulation valley for the first setting, R of Fig. 1, and at the maximum modulation peak for the second setting, R_3 of Fig. 1. The accuracy of the measurement was determined mainly by the accuracy of V_2, and was better than 1 per cent for the NADEVCEN tests.

It should be noted that the instrumentation was grounded at a point to make it compatible to the grounded neutral of the a-c input. Experience has shown that a minimum of stray interference or pickup is obtained in the instrumentation when its ground is solidly connected to the input ground.

Frequency characteristics of the voltage modulation were studied by direct oscilloscope display or photography with the use of slow sweep speeds, and also with the Brush amplifier and recorder shown in Fig. 1. The oscilloscope provided best indication when it had a d-c amplifier for the Y-axis and sweep speeds going down to 1 cps. The sweep speed was normally adjusted to approximately 5 cps. The voltage modulation was displayed on the oscilloscope as a series of sharp pulses corresponding to a peak of each half-cycle. With a slow sweep, in the order of 1 or 2 cps, the spikes blended together to form the modulation pattern. The actual modulation pattern obtained appeared to be random in nature. An example of such a pattern photographed during one sweep of the oscilloscope is shown in Fig. 2.

Table I. Results of Tests

Machine	Load, Per Cent	Modulation, Per Cent	Voltage Modulation Peak-to-Peak, Volts	Frequency Range, Cps
Generator *A*, 400 cps	0	0.80	2.60	5
	50	0.69	2.26	5
	100	0.67	2.16	5
Generator *B*, 400 cps	0	0.45	1.47	27 to 30
	50	0.34	1.01	29
	100	0.34	1.01	28 to 35
Generator *B*, 400 cps, 2 units in parallel	0	0.34	1.04	14 to 16
Generator *C*, 380 cps	0	0.38	1.20	40 to 43
	75	0.37	1.18	40 to 50
	100	0.12	0.39	40 to 50
Generator *C*, 400 cps	0	0.24	0.76	9
	75	0.37	1.18	50
	100	0.12	0.39	50
Generator *C*, 420 cps	0	0.19	0.61	52
	75	0.29	0.95	50
	100	0.18	0.58	50
Generator *D*, 400 cps	0	0.12	0.38	50
	50	0.035	0.12	50
	75	0.035	0.12	50
Generator *E*, 400 cps	0	2.26	7.35	12
	50	0.86	2.78	11
	100	0.57	1.85	11
Inverter *A*	0	0.055	0.18	7
	100	0.055	0.18	7
Inverter *B*	0	0.12	0.38	5
	100	0.12	0.38	7
Inverter *C*	0	0.11	0.37	8
	100	0.11	0.37	8
Inverter *D*	0	1.26	4.1	40 to 50
	100	1.26	4.1	40 to 50
Inverter *E*	0	0.23	0.77	45 to 60
	100	0.23	0.77	45 to 60

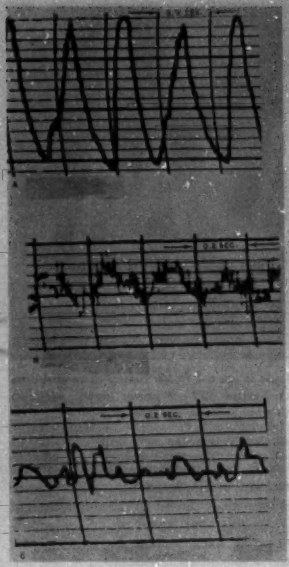
Results

The generators and inverters tested for voltage modulation are listed below:

Generator *A:* 40-kva a-c generator, 120/208 volts, 0.75 power factor, 3-phase type *A-1*, using a carbon-pile regulator and mounted on a 50-horsepower hydraulic constant-speed drive.

Generator *B:* 40-kva a-c generator, 120/208 volts, 0.75 power factor, 3-phase, type *A-1*, using a magnetic-amplifier regulator and mounted on a 50-horsepower hydraulic constant-speed drive.

Generator *C:* 40-kva a-c generator, 120/208 volts, 0.75 power factor, 3-phase type *A-1*, using a carbon-pile regulator and mounted on a 200-horsepower variable-speed electric drive.



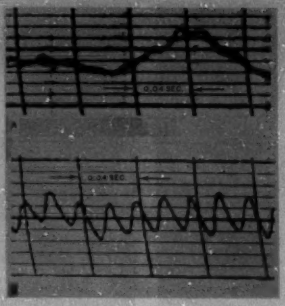


Fig. 3 (left). Oscillograms to show frequency characteristics of voltage modulation

A—Generator A, no load
B—Generator B, full load
C—Generator B, no load, two units in parallel

Fig. 4 (above). Oscillograms to show frequency characteristics of voltage modulation

A—Generator C, no load, 400 cps
B—Generator C, 75-per-cent load, 400 cps

Generator *D:* 60-kva a-c generator, 120/208 volts, 0.75 power factor, 3-phase Bureau-of-Aeronautics-type *E-1823*, 320 to 480 cps, using a magnetic-amplifier regulator and mounted on a 200-horsepower variable-speed electric drive.

Generator *E:* 9-kva a-c generator, 120/208 volts, 0.75 power factor, 3-phase Bureau-of-Aeronautics-type *E-1699-1*, using a carbon-pile regulator and mounted on a 35-horsepower variable-speed electric drive.

Inverter *A:* 2,500-volt-ampere inverter, Bureau-of-Aeronautics-type *E-1725*, with a carbon-pile voltage regulator and a carbon-pile speed regulator.

Inverter *B:* 1,500-volt-ampere inverter, Bureau-of-Aeronautics-type *E-1737-1*, with an electronically trimmed carbon-pile regulator.

Inverter *C:* 1,500-volt-ampere inverter, Specification *MIL-I-7032*,[3] with electronic regulators for close control of voltage (±2 volts) and frequency (±2 cps).

Inverter *D:* 250-volt-ampere inverter, Bureau-of-Aeronautics-type *E-1617-1*, with a carbon-pile regulator.

Inverter *E:* 750-volt-ampere inverter, Bureau-of-Aeronautics-type *E-52804-1*, with an electronically trimmed carbon-pile regulator.

The results of these tests are given in Table I.

It should be noted that, for machines normally used for 400-cps systems, the percentages of modulation were much below the 2-per-cent limit specified in various military specifications. In all cases, the values were under 1 per cent for the generators designed for 380- to 420-cps systems, and for inverters and above. This was consistent with data obtained by other engineering groups who have measured voltage modulation.

The results showed that voltage modulation was much less when a generator was driven directly from the drive stand than when it was driven from a constant-speed drive; see Table I. When the same generator was used with a carbon-pile regulator and then with a magnetic-amplifier regulator, the amplitude of the voltage modulation for all corresponding loads was much lower with the magnetic-amplifier regulator. However, the frequency of the voltage modulation was much higher. This can be observed by comparing Fig. 3(A) with 3(B). The voltage modulation was further reduced when the same generator with the magnetic-amplifier regulator was paralleled with another similar generator-regulator combination. Although not true in every case, the voltage modulation generally was reduced with loading of the generators.

Figs. 3 and 4 illustrate, in general, frequency characteristics of voltage modulation. Even though random in nature, there are one or more definitely preponderant frequencies. Even with the most random or distorted wave shape of the voltage modulation, a recurrence rate can be observed. The frequency ranges in Table I are only generalizations made from oscillograms, some of which are show in Figs. 3 and 4. In cases where the repetition rate did not remain constant, a frequency band was noted which was representative of the lowest and highest repetition rates from a number of measurements. In many cases, when more than one superimposed repetition rate was noted the most predominant frequency or frequency band was selected.

It was noted that both the low- and high-frequency components of the components of the complex modulation wave creased in amplitude as the extremes of 10 toward 2 cps and 25 toward 70 cps were approached. This was considered plausible, since, starting at 10 cps and going toward lower frequencies, voltage regulators become increasingly capable of affecting corrections and, starting at 25 cps and going toward higher frequencies, generator-regulator time constants increasingly limit any sudden voltage changes during steady-state operations.

Discussion and Conclusions

Determination of voltage modulation always involves peak-to-peak measure-

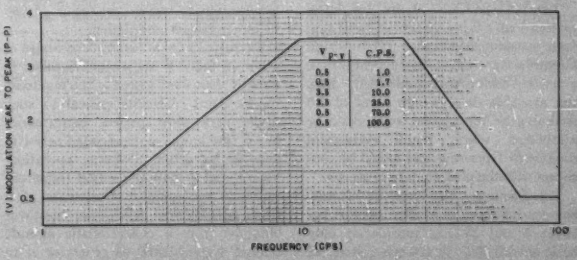


Fig. 5. Frequency characteristic limits of the voltage modulation envelope

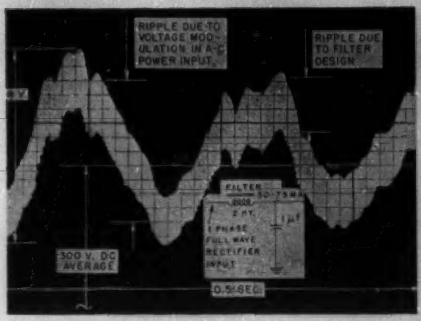

RIPPLE DUE TO VOLTAGE MODULATION IN A-C POWER INPUT
RIPPLE DUE TO FILTER DESIGN
300 V. DC AVERAGE
FILTER
1 PHASE FULL WAVE RECTIFIER INPUT
0.5 SEC.

Fig. 6. Oscillogram of ripple in the d-c power supply of an air-borne electronic equipment with 1-phase 400-cps power input, minimum operating frequency is 320 cps

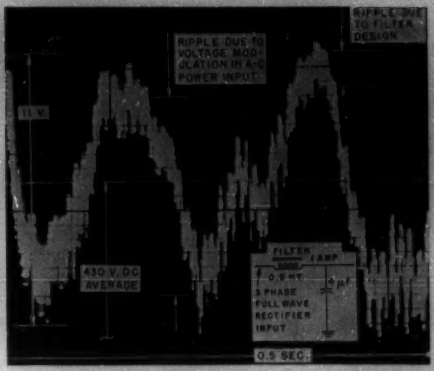

RIPPLE DUE TO VOLTAGE MODULATION IN A-C POWER INPUT
RIPPLE DUE TO FILTER DESIGN
430 V. DC AVERAGE
FILTER
3 PHASE FULL WAVE RECTIFIER INPUT
0.5 SEC.

Fig. 7. Oscillogram of ripple in the d-c power supply of an air-borne electronic equipment with 3-phase 400-cps power input, minimum operating frequency is 320 cps

ments which are converted by mathematical relationship to per-cent voltage modulation. Designers of aircraft utilization equipment, when considering voltage modulation, convert it back to peak-to-peak values for their use. Since the voltage modulation is measured and used on the same peak-to-peak basis, it would seem more desirable to define it and place specification limits on this same peak-to-peak basis. There is the possibility in the conversion of per-cent modulation to per-cent ripple that an error might be made which would reduce to one-half the actual amount of per-cent modulation which is obtained from the aircraft electric power system. This can happen when it is considered that per-cent modulation in a system is calculated as one-half the voltage swing divided by the average voltage times 100, whereas in power supplies per-cent ripple is determined[4] by the full voltage swing divided by the average times 100. Thus, 2-per-cent voltage modulation in the power input to an electronic unit power supply would provide a 4-per-cent voltage ripple in the $B+$ plate voltage supplies.

The NADEVCEN results and information from other sources indicated that voltage modulation in present a-c 400-cps aircraft power systems is much below the 2-per-cent limit presently specified. Fom the available data, a new limit of 3.5 peak-to-peak modulation volts could be established which would correspond roughly to 1 per cent and would be consistent with the present state of the art.

The frequency characteristic of voltage modulation generally ranges between 5 and 35 cps, considering only basic aircraft 380- to 420-cps systems of 1,500-volt-ampere capacity and above. When the frequency characteristic is above 35 cps or below 5 cps, its amplitude or energy content becomes proportionally less than that at the center portion. Thus, the general frequency characteristic could be defined, as shown in Fig. 5. This definition could be helpful to aircraft utilization equipment designers in designing into their equipment compatibility with aircraft electric systems.

The 400-cps powered $B+$ supplies for aircraft electronic systems, during design, must meet performance limits of ripple content. In general, the filters used are as small and lightweight as possible, consistent with the ripple limits. Thus, the cutoff frequency of the filter is usually close to but below the lowest rectified power frequency. Voltage modulation of the aircraft power system has frequency characteristics much below that of the basic power frequency and, thus, passes through the filter undiminished. It is completely impracticable to design $B+$ supply filters with a cutoff frequency low enough to reduce the ripple caused by power system modulation.

There may be a larger percentage of ripple in the $B+$ lines when installed in aircraft than when tested at the manufacturer's plant. Circuits critical to such ripple may show perfect performance under test but have depreciated performance under actual air-borne operation. Figs. 6 and 7 show the actual ripple obtained in two samples of air-borne electronic equipments. The power input was obtained from a characteristic aircraft electric system listed as generator A in the section entitled "Results." To be noted is the comparatively greater amount of ripple caused by the power system than caused by the filter design. This difference between actual ripple resulting from voltage modulation and the ripple after rectification of the a-c power anticipated from filter design alone must be considered in the design of electronic equipment.

Many types of amplifiers used in aircraft handle their signals within narrow frequency bands which lie in the frequency spectrum of 0 to 250 cps. During initial design stages the exact bandpass must be selected. If electric system designers can assure that power system voltage modulation remains in a definite frequency band within certain voltage limits, then the electronic designer can make his bandpass selections more intelligently.

Servo systems used in aircraft are susceptible to the voltage modulation in the electric power system. Autopilots have shown from experience a maximum susceptibility to voltage modulation when that voltage modulation has frequency components in the band of 0 to 10 cps, susceptibility becoming less with increasing frequency. Thus, there are direct areas of conflict between the electric power system voltage modulation characteristics and utilization equipment requirements. Any spread in the allowed frequency characteristics, as shown in Fig. 5, would create very serious problems in compatibility.

For aircraft using the 380- to 420-cps type of electric system, this conflict can be minimized by controlling electric system and electric power equipment de-

sign so that the voltage modulation frequency characteristic is within the limits of Fig. 5. Utilization equipment developments can be then accomplished with the full knowledge of these voltage modulation characteristics and can be designed accordingly.

It is concluded that electric 380- to 420-cps power equipment should be designed to that, when this equipment is integrated into the aircraft electric system, the voltage modulation characteristics should be within the following limits:

1. The amplitude should not exceed a maximum of 3.5 volts as measured peak-to-valley difference between the minimum voltage reached and the maximum voltage reached over a period of at least 1 second.

2. The frequency characteristics of the voltage modulation shall be confined within the envelope of limits shown in Fig. 5.

It is feasible to consider that presently available power equipment designs can conform to the foregoing limits and future designs should continue to maintain these limits. Thus, utilization equipment design in maintaining full and consistent cognizance of such well-defined characteristics will result in higher degrees of compatibility with aircraft electric systems.

References

1. Characteristics of Electric Power, Aircraft. *Military Aeronautical Specification MIL-E-7894*, Aero Publication and Specification Division, Naval Air Supply Depot, Philadelphia, Pa., Aug. 14, 1952.

2. General Specification for Alternating Current Generators and Regulators. *Military Specification MIL-G-6099*, Aero Publication and Specification Division, Naval Air Supply Depot, Philadelphia, Pa., Apr. 19, 1950.

3. General Specification for Inverter, Aircraft. *Military Specification MIL-I-7032B(ASG)*, Aero Publication and Specification Division, Naval Air Supply Depot, Philadelphia, Pa., Aug. 8, 1953.

4. Fundamentals of Engineering Electronics (book), William G. Dow. John Wiley & Sons, Inc., New York, N. Y., 1937, p. 508.

Discussion

Erle P. Barlow (The Glenn L. Martin Company, Baltimore, Md.): Both designers of air-borne electronic equipment and designers of air vehicle a-c electric power systems are facing equipment-versus-system-compatibility difficulties because of lack of adequate definitions for voltage and frequency modulation. The author is to be complimented for having attacked at least half of the problem.

The author's suggestion that voltage modulation be defined on a peak-to-peak basis to avoid confusion between equipment and power system engineers has merit. While it lacks the virtue of the percentage definition which automatically applies to systems of any rated voltage, the importance of avoiding misinterpretation because of the differing conventions of power producers and power users should not be overlooked. A definition is needed which will be crystal clear to both and also usable on a system of any rated voltage.

The author's investigation has been confined to modulation produced by power generation and conversion equipment. In an air vehicle power system supplying many loads, modulation also occurs because of switching transients. Of these transients, probably 99 per cent are within specified steady-state system voltage limits but their occurrence can and does induce response in certain equipments, e.g., autopilots. Frequent occurrence of such transients, even though nonperiodic, is likely to be intolerable. The same transient and resulting response occurring less frequently may constitute only a nuisance; while if it occurs very seldom, say only once or twice at some noncritical time in each flight, the situation is likely to be acceptable.

Since the factors affecting each undesired response of the equipment are magnitude of voltage variation and time rate of the variation, it is suggested that the need for an adequate definition of system voltage modulation will be satisfied only by some arrangement which takes into account all three following factors:

1. Magnitude of voltage variation.

2. Rate of voltage variation, or frequency of variation if it is of a sustained periodic nature. For an a-c system, a possible expression of this quantity for a nonperiodic disturbance would be variation of peak-to-peak voltage per cycle of system frequency.

3. Number of occurrences of the variation per interval of time.

O. Markowitz: Mr. Barlow brings to light an excellent point not covered by the paper under discussion, i.e., voltage modulation created by cyclic load variations in contrast to the voltage modulation caused and inherent with the power-producing and regulating equipment and as covered in the paper.

Should load switching force system voltage transients above steady-state limits, it can be considered within the defined scope of transient voltage limits as in present military specifications. It is considered that such switching transients are usually caused by the severe load changes and occur seldom. It is most difficult to consider or define those load variations which are random and not sufficiently severe to create voltage variations near steady-state limits.

Load modulation characteristics could only be considered when the load varies on a regular repetitive or cyclic basis. With such a cyclic load, the electric system designer has the responsibility to work out the compatibility problems between the load and its influence on the electric system characteristics. It is agreed that it would be most desirable to define the limits of such influence.

In response to Mr. Barlow's discussion, a very cursory investigation was made to look at some electric system characteristics caused by load modulation. A type-*E-1821-1* 15-kva 3-phase generator with magnetic-amplifier regulator was operated to switch a resistive load. The load was adjusted so that the voltage transients obtained were approximately at the extreme of steady-state voltage limits (about 10 volts maximum change with 3.44-kw loading per phase). The load was switched ON and OFF with intervals of approximately 1 second. The resulting transients were analyzed to see the rates of voltage change. The first cycle was not included to eliminate subtransient effects. The maximum rate recorded during the repetitive transients was during the first 4 cycles after load OFF. At the fourth cycle there was evidence of regulator correction and, beyond that number of cycles, rate of voltage change was much less. This maximum rate was approximately 1,000 volts per second. No further investigation could be made at this time to determine the validity of the 1,000-volt-per-second rate for other conditions of operation or other generator-regulator combinations.

Further study should be made to determine the basic criteria and their permitted limits of load modulation influence on aircraft electric systems. This characteristic might be best defined independently of the characteristics defined in the subject paper since the area of technical design cognizance would not be the same for each aspect.

Important Factors Influencing the Choice Between a Synchro and a Potentiometer as an Angular Position Pickoff

H. A. DINTER, JR.
ASSOCIATE MEMBER AIEE

DESPITE the present high state of perfection of both synchros and potentiometers, the aircraft industry is already designing devices the accuracy of which is better than that of the available means for transmitting output data to the point of use. The "stable platform" attitude reference and certain automatic astronomical instruments are examples of equipment presenting these new demands.

This paper does not intend to criticize the ingenious and fundamental thinking which has produced such amazing improvements in angular position pickoffs since World War II. Its purpose, rather, is to outline some of the more difficult problems which face the designer of these new systems and to indicate the more important factors which determine his choice between these two major types of pickoff. This will be done from the point of view of an organization which uses both of these types of pickoff.

The method of approach will be first to outline application factors common to all angular position pickoffs. Some of the reasons for choice of one of the two over the other and some of the problems which must be solved as a result of the choice made will then be stated. Finally, a statement of expected future needs is made, expressed in terms of existing performance limits such as a statistically significant life warranty for both types of pickoff, improved linearity and resolution for the potentiometer, and improved accuracy and application literature for the synchro. It is hoped that this approach will serve to initiate and stimulate a discussion between users and suppliers which will eventually lead to the more advanced design targets for the synchros and potentiometers of the near future.

Paper 56-449, recommended by the AIEE Air Transportation Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE South West District Meeting, Dallas, Tex., April 2-4, 1956. Manuscript submitted July 25, 1955; made available for printing February 14, 1956.

H. A. DINTER, JR., is with the Minneapolis Honeywell Regulator Company, Minneapolis, Minn.

General Problems in Applying Angular Position Pickoffs in Aircraft

Certain application problems are common to both synchros and potentiometers used as angular position pickoffs. The most important step in choosing between them is to determine what limits and what emphasis the system places on these various common problems. Probably the most common mistake made in choosing a specific type of pickoff lies in deciding at the outset that a system shall use either synchro or potentiometer pickoffs to the exclusion of the other. The system must then be designed to conform to the limits of the already chosen pickoff. The fallacy in this approach is that neither pickoff has such outstanding advantages over the other that it must be used exclusively. A better approach is to evolve a fairly complete system design before choosing the pickoffs. The pickoffs can be treated simply as electromechanical transducers, blocks in the block diagram, until the system has evolved to such a point that the important characteristics of the pickoff can be stated. The pickoff can then be chosen to meet the system requirements. Delaying the choice of the pickoffs until the system requirements have been determined should allow the system designer much needed flexibility in his difficult work. This is the approach to all of the rest of the component devices in the system, and should be the approach to the pickoffs as well. A minimum list of the system requirements which should be established follows. It is, of course, realized that, if no pickoff can meet these initially established requirements, some compensation must be made elsewhere in the system or in its specifications.

Characteristics of the Electromechanical Transfer Function

Probably the first systems requirement to be established is the desired functional relationship between mechanical input angle, and electric output signal. Various requirements for signal output from a vertical flight gyroscope might serve as examples. Within limits, the mechanical angles assumed by the gimbal shafts of this instrument are equal to the angular roll-and-pitch attitudes of the aircraft. The aileron control channel of an autopilot requires an electric signal directly proportional to roll attitude angle, so that the electric output voltage from one of the roll pickoffs should be a linear function of mechanical input angle. To prevent the aircraft from losing altitude in a co-ordinated turn, the elevator channel of the autopilot requires an "up-elevator" signal proportional to one minus the cosine of roll attitude angle, divided by the cosine of roll attitude angle. If this same vertical attitude gyroscope is to provide input signals to one type of radar antenna stabilization computer, the pickoffs on both pitch and roll axes must produce two output signals each, one proportional to the sine, and one proportional to the cosine of their mechanical input angles.[1]

The desired functional relationship between mechanical input angle and electric output signal is usually written in terms of the ratio of carrier output voltage from the pickoff to the carrier input voltage exciting the pickoff. The relationship is expressed as an equation in which this ratio is set equal to some function of mechanical input angle. The range of mechanical input angle over which this relationship must apply must also be established.

Conformance to Function

The next system requirement to be established is the error or tolerance which may be allowed between the actual and the desired electric output signals. In the examples mentioned in the foregoing, the autopilot aileron channel and the antenna stabilization computer require quite accurate conformance to the desired function, while the autopilot elevator channel can tolerate fairly large errors in the up-elevator signal. Required accuracy to the desired function is usually stated by demanding that the actual output ratio fall within a band drawn on either side of the desired function, as plotted on a graph of output ratio versus mechanical input angle. The width, at any point, of this band is usually expressed as a stated percentage of the maximum value which the output ratio reaches.

Noise

The system's sensitivity to spurious output voltages which interfere with the desired signal must be studied. Unfortu-

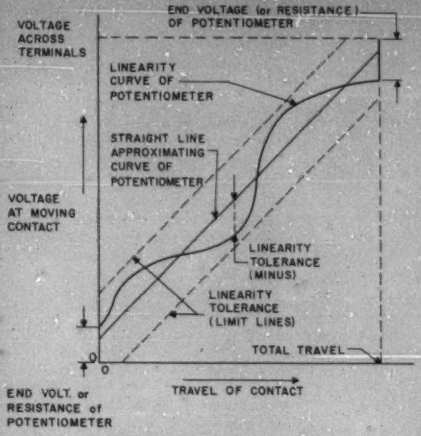


Fig. 1(A). Independent linearity

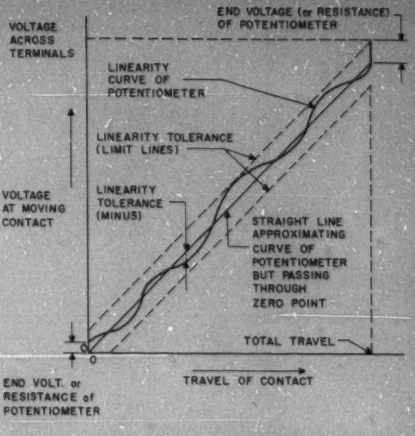

END VOLTAGE (or RESISTANCE) OF POTENTIOMETER
VOLTAGE ACROSS TERMINALS
LINEARITY CURVE OF POTENTIOMETER
LINEARITY TOLERANCE (LIMIT LINES)
VOLTAGE AT MOVING CONTACT
LINEARITY TOLERANCE (MINUS)
STRAIGHT LINE APPROXIMATING CURVE OF POTENTIOMETER BUT PASSING THROUGH ZERO POINT
TOTAL TRAVEL
0
0
END VOLT. or RESISTANCE of POTENTIOMETER
TRAVEL OF CONTACT

Fig. 1(B). Zero-based linearity

nately, direct comparison of the two types of pickoff is made difficult by three factors. First, the major sources of interference voltages are not of the random sort which can be handled by statistical means; neither pickoff noise signal approaches the lower limit set by thermal agitation noise. Second, the characteristics of the noise output voltages differ to such an extent that no uniform units, or test conditions, have been evolved. Third, the magnitudes of the various signals produced vary with application factors, and the degree to which these spurious signals may be eliminated by appropriate servo amplifier circuitry differs for the two pickoffs. As an example, resolutions of 1,000 to 1 are attainable in common synchros and potentiometers, but an improvement of 10 to 1 can be obtained by filtering and quadrature rejection for the synchro while no similar improvement can be obtained for the potentiometer.

Impedance

The electric impedances of the pickoff determine certain characteristics of the system, which may mean that one of the two types allows simpler or more efficient circuitry than the other. Among the problems related to impedance, the following are of importance: pickup of interfering signals from nearby power and signal leads, pickoff loading by other circuit elements, summing network problems, phase shift between carrier input and signal output voltages, and sensitivity of the transfer function to changes in carrier input amplitude. System requirements may dictate that the a-c or d-c carrier envelope be shifted in time phase. This operation can be performed directly only on a varying direct voltage. This can mean that those pickoffs which are usable only with an a-c carrier will make necessary a demodulator, phase-shifting network, and modulator to perform the desired phase-shifting operation.

Mechanical, Economic, and Environmental Characteristics

In addition to the characteristics of the transfer function, the mechanical reactions of the pickoff on the shaft carrying the mechanical input angle data must be considered. These torques result from bearing and other frictions, and from the electric load on the pickoff.

Designers who are disposed to use one of the two types of pickoff to the complete exclusion of the other frequently state their reasons in terms of mechanical simplicity or economy. The notion that one of the two is universally simpler to apply or more economical is simply not true. With respect to mechanical simplicity, either pickoff is available in the widely standardized "servo mounting" and, where this feature is not applicable, either may be obtained in a bearingless configuration intended for mounting, so that existing bearings on the data shaft serve also as bearings for the pickoff. With respect to expenditure, the cost of even the most expensive type of pickoff is likely to be so small a portion of the cost of the whole system that the possibility of eliminating system complexity by a wise choice of pickoffs is far more important than the cost of the pickoffs themselves.

The ability of the two types of pickoff to withstand environmental stress is not a great deal different from similar capabilities of other components. Purchasing agencies have strongly indicated that all components may soon be expected to meet far more difficult high-temperature demands than now possible.

Comparison of Synchros and Potentiometers as Angular Position Pickoffs

Only after the necessary characteristics of the required pickoffs have been established, is it wise to begin to compare the synchro to the potentiometer. It is unlikely, of course, that all requirements will be met by either of the two types; requirements should be listed with the most important first, and the ones on which compromise can be made toward the last.

The Potentiometer—Flexible Source of Complex Functions

The province of the potentiometer is at the two opposite ends of the range of complexity of the desired electromechanical transfer function. Its basic transfer function is that in which output ratio varies in direct proportion to mechanical input angle. However, the potentiometer is far simpler to modify so as to obtain complex transfer functions than is the electromagnetic pickoff. Even purely empirical functions can be obtained.

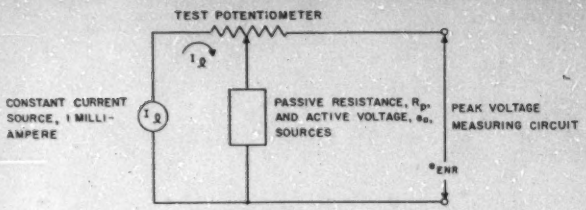


Fig. 2. Equivalent noise resistance test circuit

Reference 2 describes in detail techniques for developing complex functions. Some of the techniques listed include: continuous change in resistance per unit of angular motion by mandrel tapering and change in wire spacing; stepwise change in resistance per unit of angular motion by stepped mandrels and change in wire diameter and material; stepwise change in voltage gradient by tapping shunt resistors across portions of the resistance element, use of end resistors, use of pull-in resistors to generate reversal of slope of the function, and fixed voltage clamping applied to tap points; shunt loading of the slider; and use of cascaded, ganged potentiometers to generate power series expansions.

Another aspect of the problem of availability of the desired function is the limited range of angular motion of the potentiometer. Not even the multiturn helical potentiometer is adapted to applications in which there is unlimited angular freedom of the shaft carrying the mechanical input data. If, in a single-wiper potentiometer, the wiper is allowed unlimited angular freedom, one of two compromises must be made. If the potentiometer is wound continuously so that there is no point where the wiper contacts an insulated sector, the winding must receive its excitation voltage between two opposite tap points, the electric output function will repeat twice each revolution, and an ambiguity will result. If no ambiguity is permissible, a winding covering less than 360 degrees of the mandrel must be used. In this case, the wiper will contact the insulated sector, causing an open circuit in the summing network.

Conformity to Function

Once it has been determined that a potentiometer pickoff can simulate the desired function, the system requirements for conformity to that function must be established. Each function constitutes a special case in the matter of conformity, so that only the more limited case of the linear function will be considered. So much data have been published with respect to this subject,[3-7] that a mere listing of the definitions and their descriptive figures, drawn from reference 5, should suffice. The figures from reference 5 are reproduced as Fig. 1.

1. Independent linearity is the maximum deviation of the actual electric output expressed as a per cent of the total electric output at any point from the best straight line drawn through the output versus displacement curve. (This line shall be measured throughout the effective electrical travel.) The slope and position of the straight line from which the linearity deviations are measured must be so adjusted as to minimize these deviations.

2. Zero-based linearity is the maximum deviation from the best straight line passing through zero voltage at zero shaft displacement.

3. Terminal linearity is the maximum deviation from a straight line which passes through the zero and 100-per-cent voltage points based on the effective electrical travel.

Other definitions, similar in result, are to be found in the noted references.

Linearity, stated in per cent of the maximum value of output ratio, is really quite meaningless as a means of comparing potentiometer pickoffs as to relative quality. If, for example, two potentiometers, one having ten turns and the other one turn, but of equal diameter and winding quality, are compared in terms of per cent of maximum ratio, the 10-turn potentiometer will falsely appear to be ten times more linear than the 1-turn potentiometer. A far better figure of merit for potentiometers used as pickoffs may be obtained by converting the percentage linearity figure first to degrees linearity and then to inches along the wiper travel circumference. In this way, the quality of the resistance elements may be compared independently, apart from the confusing factors of the mechanical mounting of which they may be a part

$$\Delta_\theta = (\Delta_p/100)\theta_t$$

where

Δ_θ = error, degrees
Δ_p = error, per cent
θ_t = total winding length, degrees

$$\Delta_l = \Delta_\theta d(0.00872)$$

where

Δ_l = error, inches
d = winding diameter, inches

It must, of course, be realized that linearity error stated in inches includes errors caused by lack of uniformity in the wire and mandrel as well as actual misplacement of turns, but linearity in inches does provide a fair and reasonable figure of merit for comparison of potentiometer quality.

The best modern potentiometers exhibit a tolerance of ±0.007 to 0.010 inch along the wiper track. This is equivalent, on a 3-inch-diameter bobbin, to a tolerance of ±0.3 to 0.4 degree. Average potentiometers have a tolerance of about 12 mils, which is equivalent to ±0.5 degree.

Accuracy in degrees is, itself, a figure of much interest. Only occasionally is a pickoff application met in which the input angular information occupies more than 360 degrees of shaft motion. Frequently, all of the information is contained in a range far smaller than 360 degrees. Unless some means, usually gearing, is used to amplify the shaft motion so as to use all of the pickoff winding length, the stated percentage tolerance simply does not apply. If the pickoff shaft is directly coupled to the mechanical input data shaft, the linearity in degrees applies.

Basically, then, the potentiometer is best suited to applications in which shaft motion is limited to less than 360 degrees of arc, and which require no better than 0.3- to 0.5-degree accuracy. In any case, it should be stated that a lower limit is set by the width of the winding wire. Conformity cannot be better than one-half the degree width of one wire or, expressed in another way, one-half the percentage of the total resistance which is included in one turn if the total winding length is used.

Two Means of Stating Noise Figures

There is a considerable area of disagreement between the potentiometer manufacturers as to what effects should be included as voltages which interfere with the desired output signal, and as to what units should be used to represent their magnitudes.

I. J. Hogan lists a number of sources of interference in an unpublished paper. To each source he assigns a particular unit of quantity and states what external application factor determines the size of the

interfering voltage produced by that source. An example of particular interest is the two components of resolution noise; one is caused by the winding current, the application factor, flowing through the resistance of a single turn, which is the noise figure unit chosen, and the other is caused by the load current, the application factor, flowing through the fluctuating contact resistance, which is the unit chosen for a noise figure.

J. R. Altieri[8] does not separate the various sources of interference voltages but combines them all into a single equivalent noise resistance. This he does by establishing a standard measurement condition requiring a load current of 1 milliampere. His proposed circuit is indicated in Fig. 2. The equivalent noise resistance relationship then becomes

$$ENR = e_{ENR}/I_1 = (e_a/I_1) + R_p$$

where

ENR = equivalent noise resistance
I_1 = standard 1-milliampere load current
e_a = active noise voltage
R_p = passive noise resistance

Both systems seem to have considerable merit. Altieri's system would be especially adapted to inspection and specification applications. One point must be observed however; the test method does not reveal the full magnitude of the resolution step voltage, detecting only that part of this voltage caused by flow of load current through the fluctuating contact resistance. It thus can minimize one of the major potentiometer problems. Hogan's method, if it can be developed along the lines suggested in his article, would be of greatest interest to the designer, for only by having the noise sources separated and classified according to the effect of particular application factors on their magnitudes will it be possible to predict noise in any specific application.

To date, only one manufacturer has published noise data; this is in terms of equivalent noise resistance. The value stated for its products is 140 ohms. This area demands clarification, as interference problems have become most demanding. In one modern high-performance fighter aircraft, controlled elevator movements of an amplitude not in excess of ±0.016 degree would be required to maintain altitude within a 200-foot range.

High Generator Impedance

In general, a lower limit must be set upon end-to-end potentiometer resistance to obtain good linearity and resolution. This varies from 1 to 10,000 ohms. This relatively high source impedance creates several problems, the most important being the loading problem. A loading resistor in parallel with the output voltage terminals distorts the output ratio. An excellent chart showing the magnitude of this effect is given in reference 9. Several methods of compensating for this effect are also suggested.

When a series voltage summing network is used, the high source impedances of the individual potentiometers become a problem. Most of the summing network operates at a high impedance to ground which may cause hum pickup and excessive flow of signal currents through lead-to-ground capacitances.

A marked advantage of the potentiometer is that it may be operated with a d-c carrier voltage. If system stability requirements necessitate addition of lead networks, these may operate directly on the output signal without modulators and demodulators.

Mechanical Loading

The electric loading of the potentiometer is not reflected back into the shaft carrying the mechanical input angle. Reflected torques are, rather, frictional, arising in the bearings and wiper. Their magnitude varies with wiper arm radius and with required wiper tension. Values of starting torque less than 1 inch-ounce can be obtained in 3-inch-diameter single-turn potentiometers with integral bearings; with a few exceptions, these torques are negligible.

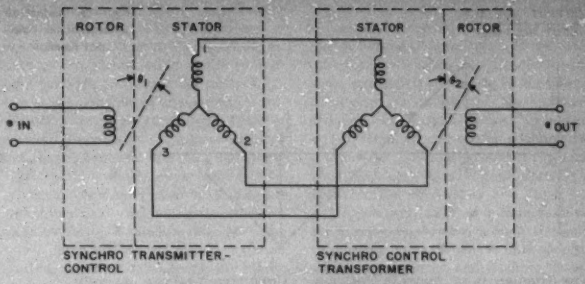


Fig. 3. Synchro connection schematic

In applications such as gyroscope pick-offs, even this torque may be too high. In this case, and in others where design flexibility demands, the easily separable wiper and resistance element may be mounted so as to make use of the bearings which carry the input data shaft. The wiper assembly can be made quite light and compact so as not to present large bending loads to the overhung shafts upon which it must be mounted.

Availability of Application Literature

Before considering the synchro, it is necessary to note one final point which often dictates choice of a potentiometer over a synchro. This matter is the availability of application literature. The potentiometer manufacturers make easily available a mass of such literature, extensive handbooks, reprints of pertinent technical articles, very full bibliographies, and monthly house organs containing recent data. Until the appearance of reference 3, the synchro manufacturers had done almost nothing in this field. The availability of applications literature of good quality, especially containing solved examples, is an important factor despite not being a characteristic of the pickoff itself. A realistic appraisal of just what data with respect to synchros are really proprietary should make it possible for the synchro manufacturers to release much application data not available now.

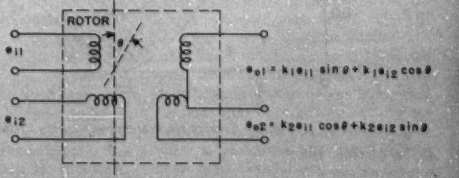


Fig. 4. Synchro resolver

e_{11}, e_{12}—Input voltages
e_{o1}, e_{o2}—Output voltages
k_1, k_2—Transformation ratios
θ—Shaft angle

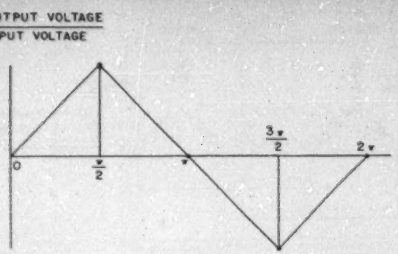


Fig. 5. Induction potentiometer output function

The Synchro as an Angular Position Pickoff

The first device to which the now widely used term synchro was applied was not really intended primarily as an electromechanical transducer. It was, rather, the still widely used remote shaft position indicating system. These synchros are connected in pairs, consisting of a torque transmitter, and a torque receiver, or in longer chains, with the basic transmitter and receiver separated by one or more torque differential synchros. Since both input and output from these chains are in the form of mechanical angle, for a system which is intended solely to transmit angular data from one place to another in a convenient manner without auxiliary equipment and with an accuracy in the order of 1 degree, the torque transmitter-receiver pair is undeniably the most economical and reliable means available, with the obvious exception of a short, solid shaft.

The basic shaft angle to voltage-output voltage-input ratio relationship for a synchro is: voltage output/voltage input equals a constant times the sine of the shaft angle.

Unlimited Angular Freedom

A major advantage of this system over the potentiometer pickoff lies in the fact that there is no limit to the angular range of the data shaft. The output function will be repeated without any ambiguity once each revolution of the input shaft, and without limit to the number of input shaft revolutions. The term synchro has now been applied to several different devices similar in construction to the original but of differing accuracies and uses. Those to be considered here are:

1. Three-winding synchros.
2. Synchro resolvers.
3. Induction potentiometers.

The three winding synchros are normally connected in pairs, as indicated in Fig. 3, consisting of a synchro control transmitter and a synchro control transformer. This pair of pickoffs actually forms a complete summing network producing a function of the form: voltage output/voltage input equals a constant times the sine of the difference of the two shaft angles $\theta_1 - \theta_2$.

This combination is ideally applicable where the voltage output from the second synchro is used to excite an amplifier which operates a motor which, in turn, drives the shaft of the second synchro to a position where output voltage is zero. At this point, the two shaft positions are the same within the system error limits. This system is, however, frequently applied where the output voltage from the second synchro is summed with other similar output voltages and the sum voltage is made zero by a servo system. The errors attendant upon summing voltages which are proportional to the sine of shaft angle rather than directly proportional to the shaft angle are small as long as the shaft angles themselves are small, or as long as each output voltage is matched by another of about the same magnitude but of opposite phasing. It might be worth noting, before considering the synchro resolver, that the three output voltages from the stator of the synchro control transmitter of Fig. 2 bear the following relations to the exciting voltage:

Voltage output 13/voltage input = (constant) times the sine of (shaft angle).
Voltage output 32/voltage input = (constant) times the sine of (shaft angle + 120 degrees).
Voltage output 21/voltage input = (constant) times the sine of (shaft angle + 240 degrees).

Thus, the synchro control transmitter alone might be considered to be an angular position pickoff without connection to a synchro control transformer.

The Resolver

The synchro resolver, like the 3-wire synchros, has, basically, a sinusoidal relationship between voltage ratio output and shaft angle input. However, the resolver has two windings on the stator, situated 90 degrees mechanically from each other, and two windings on the rotor, also situated 90 mechanical degrees apart, as illustrated in Fig. 4. The resolver is especially adapted to systems in which the pickoff voltage output is one of the inputs to an analogue computer, such as the antenna stabilization system mentioned in the beginning of this paper.

The Induction Potentiometer

The induction potentiometer produces a transfer function such that the ratio of the voltage output/voltage input varies directly with shaft angle input rather than sinusoidally, as is the case with the other basic synchro designs. The basic period of this function, however, is one-half revolution of the input shaft so that, unless shaft rotation is limited to less than 180 degrees, an ambiguity exists, as indicated in Fig. 5.

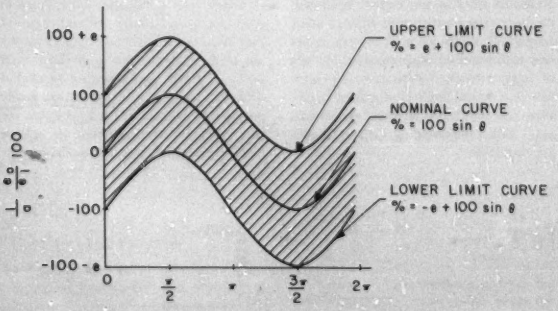


Fig. 6. Synchro resolver output accuracy limits

e_r—Transformation ratio
e_o—Voltage output
e_i—Voltage input
e—Error, per cent

Accuracy to Function

The accuracy to function for the first two of the three electromagnetic devices considered is unquestionably better than for any potentiometer pickoff. For the 3-wire type of synchro, the error for units of size 11, i.e., of 1.1-inch diameter, is 6 minutes maximum. This 0.1-degree error for an average unit compares with the 0.5-degree error for the average potentiometer. Special units may be obtained having errors not in excess of 2 minutes or 0.03 degree as compared to the special unit accuracies of possibly 0.3 degree in a potentiometer of reasonable (3-inch-diameter) size. These accuracies are a maximum deviation from a zero error assumed at electrical zero.[3]

For resolvers which are to be used for their primary purpose, trigonometric computation, tolerances are usually stated in terms of per cent of maximum coupling voltage. An accuracy stated in degrees becomes meaningless at the maximum points of the sine and cosine curves. This means of stating accuracy is indicated in Fig. 6. An idea of the accuracy of the device on a basis common with that established for the other pickoffs may, however, be obtained by reducing stated accuracy in per cent to its equivalent in degrees at the zero amplitude point of the nominal sine or cosine curve. The usual accuracy stated for a wide range of these devices is of the order of ±0.05 per cent of voltage at maximum coupling. This is equivalent, at 0 degree on a sine curve, to less than ±2 minutes of arc. The ±0.05-per-cent accuracy advertised can be compared directly to advertised accuracy, conformity to function, of a typical potentiometer designed specifically to produce a sine-cosine output. The accuracy of the potentiometer is ±0.5 per cent, only 1/10 as good as that of the electromagnetic pickoff.

Functional accuracy of the induction potentiometer is also usually stated in per cent of the maximum output voltage. For a typical design this is ±2 per cent or about 3.6 degrees at the worst points, and varying down to 0.5 degree near zero.

Noise Voltages

The content of the spurious output signal produced by a synchro differs considerably from the content of the spurious output signal produced by a potentiometer. A major contribution to the spurious output of a potentiometer is the resolution step in voltage produced as the wiper moves from one wire to the adjacent wire. The synchro class of pickoffs are essentially transformers having continuously variable coupling. They, therefore, do not suffer from stepwise voltage changes. The two important components of synchro spurious output voltage are quadrature and harmonics of the excitation frequency. As voltage output from a synchro decreases from its maximum coupling value toward zero, its time phase shifts, at first slowly, then quite rapidly within a degree or so of null. If the output is then examined by a phase-sensitive detector furnished with a reference voltage in phase with the voltage at maximum output, the in-phase component of fundamental frequency will be observed to reach zero. The angle at which this occurs is defined as one of the two nulls for the synchro. That portion of the remaining voltage which is of fundamental frequency, but 90 degrees displaced in time phase from the voltage at maximum coupling, is termed quadrature. Reference 3 states a "rule of thumb—frequently used" for the relative magnitude of this quadrature signal: "The fundamental quadrature null of the synchro shall be less than one-half the permissible angular error translated into voltage."

The other major portion of the remaining voltage consists of time harmonics of the fundamental excitation voltage. Reference 3 states that this and other components of the null are permitted to be of such value that the total null voltage, measured by an instrument having rms response, is from 1.5 to 1.6 times the quadrature voltage permitted. The designer needs to know both the rms and the fundamental null voltages so that he can design a filtering and phase-discriminating amplifier that will control his servomotor and will predict its efficiency.

Low Source Impedance

The major impedance problems of the synchro-type pickoff have to do with the fact that it is an inductive-type device with ferrous core material. The air gap is relatively large so that leakage inductances are larger than in a static transformer. Output voltages lead input voltages by 5 to 10 degrees times phase, which means that the designer must either provide for similar units, used in equal numbers in each of the chains which separate the common excitation source from the common summing point, or shift the time phase of the output from each chain to a common value prior to introducing it into the summing network. Another important consideration is the excitation voltage rating. In a potentiometer, the lower limit to excitation voltage is determined only by the size of the generated electromotive forces produced in the winding-wiper contact, and the upper limit is set largely by the power dissipation capabilities of the resistive element. In the synchro, both the upper and lower limits are set by nonlinearity of the iron. For best conformity and good null, excitation voltage should be held near the rated value.

The relatively low output impedance of the synchro-type pickoff is likely to result in a system less sensitive to hum pickup, wiring capacitances, and loading. Summing is usually done in a series string. This is especially easy since the stator and rotor of each pickoff are electrically isolated. The long leads from remote points can be grounded at one point, and the rotor output leads from the final synchro in each chain, which are connected in series to perform the summing function, can be kept short and near the amplifier input. Loading does, however, become a problem where the very highest accuracy is desired. Resolvers in computer circuits are, for example, almost always operated with compensating, or booster amplifiers which present a high impedance to the excitation supply, a low impedance to the resolver itself, and, in addition, correct the complex transformation ratio of the resolver at maximum coupling to 1.0 phase angle, zero degrees.

One point, sometimes a disadvantage, has already been mentioned. The synchro is exclusively an a-c carrier device. If phase shift in the modulation envelope of its output is necessitated by system stability demands, the most practical means of obtaining it is to demodulate the carrier, perform the desired operation on the resulting direct current, and remodulate the carrier with the modified information. The effect of this may be to add many extra components in a critically small system.

Electromagnetic Loading

Exclusive of the bearings and the slip rings, which are common to both types of pickoff, the synchro presents no coulomb friction to the data shaft. Its torques are of the springlike variety, applying centering forces to the data shaft. Unfortunately, not much work on this point has been published by the synchro manufacturers, but indications from gyroscope work are that these torques vary with shaft position from zero to a maximum of the same order of magnitude as the wiper friction torques of a potentiometer. New work or publication of existing work is badly needed in this area.

Recommendations for the Future

One matter common to both types of pickoff demands mention prior to the suggestions concerning requirements specific to each type of pickoff. This important matter has to do with the commonly used, and only poorly defined term, reliability. The equipment manufacturer's own customers are beginning to demand that he furnish some kind of life period guarantee. He can hardly do this if the manufacturers of the synchros and potentiometers carefully avoid all mention of the matter. In regard to vacuum tubes he has demanded, and obtained, such assurances. It is almost certain that the reliability of synchros and potentiometers, already frequently criticized, will become a vitally demanding problem. Many reasons for not doing anything about the problem can be proposed. Some of these are: cost of items tested to destruction, lack of knowledge on the part of the manufacturer as to what environment his product meets in the user's equipment, lack of an available means of measuring actual time of use so as to check life against specified life, and lack, of means to obtain and evaluate failed units and failure data. The most important problem, however, seems to be lack of recognition of the problem. The reply to all the others is that the tube manufacturers met these problems, and many others unstated here, and are in process of solving them. It is then suggested that a small board of interested and competent mathematicians, engineers, statisticians, and user representatives be established to consider the problem.

The Potentiometer

It is to be hoped that the improvements suggested will soon come into being. An improvement of 5 to 1 in angular linearity to about ± 0.1 degree would open up immense new possibilities to the potentiometer. Even now, autopilot applications are forcing a new evaluation of inductive pickoffs in areas where formerly potentiometers were accepted as a matter of course. Another area in which quick help is demanded is the matter of stepwise resolution. Some present-day servocontrol loop gains are such that the oscillation caused by a wiper hunting between two wires has destroyed valuable equipment. Also some new means of decreas- wiper pressure is needed while maintaining reliable contact so as to reduce both wear and input shaft friction.

The Synchro

Accuracy of the 3-wire synchro would become about as good as the accuracy of evaluation and application circuitry if the presently very special ± 1- and 2-minute synchros were made generally available. The situation with respect to resolvers is considerably darker. It has been said that, unless some means of obtaining resolvers to an accuracy of ± 0.005 per cent of maximum coupling became available, important new navigation systems would simply never work. This is a 10-to-1 improvement and demands as much work toward a means of testing such new devices as it requires for the development of the devices themselves. Once again, the need for application literature is most demanding.

It is sincerely hoped that this paper will provide a few guide points for the systems designer faced with the choice of a specific pickoff to meet established system requirements. It is almost desperately hoped that it will initiate a wider discussion of industry's needs and of the pickoff supplier's ideas for meeting them.

References

1. Radar Scanners and Radomes (book), W. M. Cady, M. B. Karelitz, L. A. Turner. McGraw-Hill Book Company, Inc., New York, N. Y., 1948.

2. Potentiometer Handbook. Technology Instrument Corporation, Acton, Mass., 1954.

3. Synchros for High Performance Servo Systems. Kearfott Company, Inc., New York, N. Y., 1955.

4. Accuracy of Potentiometer Linearity Measurements, Robert McDonald, Irving Hogan. *Tele-Tech and Electronic Industries*, New York, N. Y., vol. 12, 1953, pp. 88–90.

5. Potentiometer Definitions. *Helinews*, Helipot Division, Beckman Industries, South Pasadena, Calif., no. 2, Mar. 1953, p. 2.

6. *Helinews*, no. 8, May 1955, pp. 1–2.

7. Production Test Method for the Determination of the Linearity of Precision Potentiometers, Joseph R. Altieri. *Laboratory Report No. 1, 2, 3*, Technology Instrument Corporation, New York, N. Y., pp. 7–8.

8. Precision Potentiometers and Noise, Joseph R. Altieri. *Laboratory Report No. 6*, Technology Instrument Corporation, Acton, Mass., pp. 23–26.

9. Computing with Servo-Driven Potentiometers, F. R. Bradley, R. D. McCoy. *Tele-Tech and Electronic Industries*, New York, N. Y., vol. 11, 1952, pp. 95–97.

10. Resistors: Variable, Precision. *NAS 710*, Aircraft Industries Association of America, New York, N. Y., 1955.

Discussion

W. L. Deniston (Fairchild Controls Corporation, Los Angeles, Calif.): One of the ideas behind this paper, that of stimulating discussion between the user and the supplier, is excellent. The supplier always feels that he doesn't know enough about his customers' business. We are always eager to help the user, or potential user, to apply our products and also to produce special products to meet his special needs. One of our main sources of trouble is the lack of understanding and communication between ourselves, as the supplier, and our customers, as the users of our products.

To help overcome this problem, we have undertaken a very extensive technical and application-type information program. In addition to this, we are always ready to try to solve specific problems by special techniques which may not be applicable to standard products.

The author has recognized one of the major advantages of potentiometers. The potentiometer can supply both linear and nonlinear functions to very great accuracy. Many clever systems designers specify nonlinear potentiometers to compensate for other nonlinear elements in the system. The author's example, in which he wants to produce a particular trigonometric relationship, is also typical of many uses. We are continuously supplying nonlinear potentiometers to specific empirical functions furnished in either tabular or graphic form.

The figure of 1,000 total turns (indicating the resolutions) of a potentiometer in the 1,000- to 10,000-ohm resistance range is rather typical for single-turn units. However, we do have a number of standard single-turn models in which the 10,000-ohm resolution is better than 2,000 turns, total. We have produced, on special order, larger diameter potentiometers the resolution of while, in this resistance range, would be more than twice these values, or better than 4,000 turns. In addition to this, we can, at considerable increase in cost, use special wire alloys which will give still more resolution, perhaps as much as two or more times. This is typical of what can be done on a special basis. In this case, the resolution could be improved by almost an order of magnitude over the figure given, if it is important enough to warrant an increase in size and cost. The problems of pickup and phase shift of the output signal are very easily avoided by the use of direct current as the exciting voltage.

There are several approaches to reducing the loading error (see reference 10 of the paper). It is also possible, in many cases, to design the compensation into the winding of the potentiometer if the use conditions are known to the manufacturer.

As far as potentiometers are concerned, the statement that the "sensitivity of the transfer function to changes in carrier input amplitude" is similar to the requirement in selsyns for a common reference a-c supply to the selsyn system, in that, for accurate operation, the potentiometer input voltage must serve as a common reference for the receiver

(of the potentiometer output). The transfer function for the potentiometer is affected by changes in the potentiometer's input carrier amplitude as in the case of a selsyn system's sensitivity to supply phase. Since these are analogue devices, the output bears a direct relationship to the input which means that, if the input amplitude is reduced to half, the output amplitude is also reduced by the same proportion. If this exact relationship is not maintained, the device is useless.

The author's suggestion to state potentiometer linearity as an error in inches of wiper travel along the coil seems to be a very poor thing from the user's point of view. The manufacturer sometimes is interested in comparing coil quality in this fashion, but for the user such a figure would be absolutely meaningless as such. He would have to take this figure and relate it, in some fashion, to its effect on the output. After all, he wants to know the maximum inaccuracy of the output at any point, and this is exactly the information he has with the present potentiometer linearity definitions. If the coils in a single-turn and a 10-turn potentiometer are of the same quality and the mechanical accuracy of the assemblies is of the same order of magnitude, the 10-turn unit really does give ten times more accurate results. This is not a false appearance, as the author states in his paper. It is completely true and is exactly what would be measured by anyone checking the output accuracy of the unit.

We have, in special cases, stated accuracy as a percentage of the total travel. Under some unusual conditions, this is the information desired by the user. But in the vast majority of the cases, the user is interested only in knowing how close the actual output comes to the desired output at any place in the travel. As the author states, if only a portion of the winding is used, an accuracy figure based on the total winding length does not give directly useful information. The best answer is to measure the output accuracy over the portion of the winding in question. We frequently do this if the active length is specified.

As far as potentiometer life is concerned, my company has tested thousands of units to destruction and has accumulated a vast amount of information on the subject of potentiometer life. Even so, we do not feel that we have enough information to make a life guarantee for any potentiometer or group of potentiometers without knowing all of the parameters of the intended use.

One possible answer to this problem is to have the supplier run life tests under laboratory simulated use conditions, being certain to use a large enough sample so that the results are statistically valid. This is the approach that we have taken many times and are willing to do for any customer that wants life information for his specific application.

We are certain that, if the users and the suppliers were to get together on some sort of committee or panel basis to set up test procedures or criteria for failure, satisfactory life definitions could be worked out and the suppliers would then be willing to give life guarantees under these standard test conditions. We would be willing to co-operate with any such group which could be formed.

The best answers potentiometer suppliers can presently give to the problems of accuracy and resolution are to go to the multiturn potentiometers. This will give accuracies of ±0.025 per cent on a standard basis, and values three or four times that good can be obtained with very special units. The resolution of our standard 15-turn 10,000-ohm potentiometers is 20,000 turns of wire, or 0.005 per cent per turn. This figure can be increased by a factor of 3 or 4 at a considerable cost increase. A great deal of work is being done in the infinite resolution type of potentiometers. This should provide a partial answer to the resolution problem but, as presently supplied, these units do not have accuracy as great as the comparable wirewound units, by a factor of about 5 or 10. This situation will improve some as experience is gained in producing these units. However, there is no reason to believe, at present, that they will ever be as accurate as the existing wirewound units.

Much work is also being directed toward the problems of low-noise contacts, which give long reliable life, and it is reasonable to expect that considerable improvement will be evidenced in this area in the near future. It is certainly an excellent idea for any component user to state his future requirements so that the supplier may study them and concentrate his research efforts in the areas that will produce the desired results.

Arnold E. Hayes (Clifton Precision Products Company, Inc., Clifton Heights, Pa.): The paper presented by Mr. Dinter does a service to the users and producers of both potentiometers and synchros. The proposed and hoped for wide discussion of industry problems will inevitably expedite their solutions. While this paper avowedly advances the point of view of systems design engineers, these comments will be advanced from the experience and problems of synchro manufacturing.

The problem of accuracy and error has been rather vigorously attacked in the last few years by both synchro and potentiometer manufacturers. There has been a rather steady stream of improvements and the continuing prospects are bright:

1. Computing resolvers compensated to ±1 minute of arc and 0.03-per-cent functional error are obtainable.
2. Multipole resolvers compensated to half of these tolerances have been made.
3. The pressure of weight and space has prodded synchro development to the point where the real small units (0.750-inch outside-diameter by 1.240 inches long) today in regular production carry error limits which a decade ago were considered a model shop matter for units of much greater size.

Synchros, and potentiometers too, can certainly be misapplied. Any asymmetrical electromagnetic loading of a synchro will cause a measurable error in heading and produce reaction forces on the prime mover or transducer.

Asymmetrical loading of a pancake gyro synchro results in a precessional force on the gyro, and the stiffer the synchro, or lower the impedance, the greater is the reaction force the unit is capable of generating. While asymmetrical loading of gyro pancakes to produce a centering torque has been deliberate in some gyro designs, this directional preference can be made to approach zero as closely as the designer wishes or can detect. It need not necessarily be the same order of torque as that caused by the wiper friction of fine gyro potentiometers.

The contact problem is not a major one for synchro manufacturers. Synchro salesmen will argue that the small diameter of their highly concentric slip rings permits real contact pressure between noble metal alloys, giving the advantage of long life and less wear under actual operating conditions and environmental testing.

The problem of guaranteeing service life is very difficult for the component manufacturer. A synchro properly chosen and properly handled, installed, and protected in control equipment will last the proverbial lifetime and generally outlive other components and parts in the equipment. Careless handling and exposure to needless damage—there is no real adequate test for the human environment problem—are life-shortening factors much beyond the control of the synchro manufacturer.

Information feedback on field service problems of synchro users is, as Mr. Dinter states, very rare. Closer co-operation and information on field failures would probably uncover presently unsuspected opportunities for improvements.

Unless the system is in really deep operational trouble, however, we note a remarkable reluctance on the part of system design engineers to make or incorporate improvements, and often forego them rather than make a change.

The suggestion made by Mr. Dinter about more comprehensive application literature, including solved examples, is well taken. As in the past, the problem is not a matter of proprietary data; it is a matter of time and talent. The possible combinations are astronomical in number; editors will have to find a balance between merely mechanical and electrical differences and tabulate an enormous amount of existant data.

Finally, and it may be a consequence of industry competition, we note that the synchro companies have become much more agreeable lately to nonstandard units that help optimize the customer's system and fit his particular service requirements.

Harold Graifer (ARCA Division, Beckman Instruments, Inc., South Pasadena, Calif.): While we are in general agreement with the author's contentions, certain objections are in order. These will be stated in the order in which they arise in the paper.

We prefer that, in the discussion of noise, the resolution of a synchro be stated as infinite. Those noise voltages which arise are by definition quadrature and harmonics. Since a synchro pickoff is always used in conjunction with the natural filter and phase-sensitive detector characteristics of a 2-phase motor, quadrature and harmonics do not enter into the accuracy of resolution. It is granted that this noise would tend to produce heating in the associated components and saturate the amplifier. The limiting factor on the resolution obtainable is caused by the associated mechanical components. However, this is equally true of the synchro and the potentiometer.

In three instances the author infers that, where system stabilization requires lead

networks, the logical choice involves the use of a potentiometer as opposed to a synchro since this shifting is more readily carried out in a d-c system. The general inference is that, for similar shifting in an a-c system, the complexity of demodulation, phase shifter, and modulator lessens the appeal of a synchro. We would like to state that, while designers of the past have employed this method of information shift, it is possible with newly developed notching networks and inertial-damped motors to overcome most system stability problems where the tachometric feedback will not provide sufficient amounts of correction.

The author's dissatisfaction with the amount of application literature available, while somewhat justified, is not entirely correct in that most of the manufacturers have made considerable information available by means of articles published in trade magazines. *Control Engineering*, *Product Engineering*, and *Electrical Manufacturing* are notable examples. In many instances the application of a synchro to a system can not be handled in a routinized fashion and the indiscriminate use of application literature may result in further problems for the synchro manufacturer. We have always been most willing to assist directly any user in the application of a synchro.

It should be stated that, where a-c systems are used, potentiometers (particularly multiturn) will not of themselves prevent phase shift. Both synchros and potentiometers will create shift which may require correction for stabilization.

H. A. Dinter: The discussions presented by Mr. Deniston, Mr. Hayes, and Mr. Graifer were very much appreciated, especially because they make available precisely the type of information which the paper requested. Continued improvement of the product is rapidly making available pickoffs having the desired performance.

The point concerning a figure of merit for potentiometer linearities merits not so much a defense as an explanation of my viewpoint. The application considered was to the field of gyroscope pickoffs. In this case, all of the information which the gimbal shaft can deliver is contained in one revolution. Unfortunately, the torque required to drive a 10-to-1 gear train is simply not available from the instrument, making the use of a one turn pickoff imperative. In such an application as this, the figure of merit stated is a pertinent one. While this application is rather specialized, it is not unique. It is one of the large class of pickoff applications wherein the energy necessary to actuate the sensor or transducer, and its pickoff, must be derived from the phenomena being sensed. Even in the case of pickoff applications in which an unlimited drive torque is available, there remains the problem of finding a gear train of an accuracy an order of magnitude better than that of the multiturn potentiometer which it drives.

Analyzing Missile Electric System Reliability

B. J. WILSON
ASSOCIATE MEMBER AIEE

Synopsis: The performance reliability of a missile electric system may be defined in terms of steady-state values of voltage and frequency applied to the system equipment. This viewpoint stresses the auxiliary function of furnishing electric power as contrasted to the prime functions of electronic equipment such as signal amplification or control response. Other failure-producing parameters such as operational environments, aging, and production differences must be recognized by the definition. Hypothetically at least, these parameters may be classified as secondary factors, thus bringing attention to focus on problems most acute in early design stages. Performance reliability of load equipment is evaluated at the so-called blackbox level. In this process, the prime electric source is connected to the input; performance is measured at the output. The test results from varying input conditions furnish information to describe the input-output relationships. In the discussion, variations on electric input conditions are confined to voltage but may be extended to frequency as well by an identical procedure. From these results, statistical calculations derive values of probable performance for equipments. A similar process gives probable performance of generators or other sources, the output in this case being electric voltage. These tests replace conventional circuit analysis and testing, and account for performance dispersion that cannot be predicted with accuracy in equipment of great complexity. Hence, approximations of performance expected from subsequent production-line equipment may be obtained on breadboard equipment that cannot be submitted to environmental testing. To carry out the method thoroughly requires a single prototype model for each of the equipments that make up the entire system, or at least those that are considered most critical. Such tests do not consider explicitly the individual component parts which make up the equipment package. They are designed to permit determination of performance reliability of over-all complex systems with the least amount of system or equipment breakdown. Finally, under assumptions of statistical independence, equipment reliabilities are serially multiplied to determine over-all system reliability.

Statement of the Problem

THE reliability of an electric system is a measure of its ability to furnish the missile and its equipment the support services required of it. These services generally consist of supplying voltage, frequency, and power within certain limits. But it is impossible to determine these limits without considering the quantitative output performance of equipments as a function of voltage and/or frequency. So the scope of the missile electric system reliability problem is necessarily extended to include consideration of the output performance of system loads as functions of steady-state voltage and frequency.

Customarily, reliability is defined as the probability of performing a specific task subject to a given variable, parameter, and/or environmental values for given intervals of time. As considered here, typical tasks may include furnishing specified motor torque or speed, amplifier gain, or receiver sensitivity. Among the variables are steady-state values of voltage and frequency; among the parameters are vibration and temperature. Steady-state values are selected as the characteristic measure of voltage or frequency most common among systems and generally most in use to stipulate a desired voltage or frequency condition. Incidentally, the reliability of furnishing electric power is being defined here in distinction to the reliability of response to guidance and control signals under ideal electric power conditions.

Broadly speaking, reliability may be measured on the basis of past performance when the number of times the system and its equipments have functioned successfully is ratioed to the total number of times the system has operated. This ratio roughly describes how successful a specific system has been. It is often a useful criterion on which to judge how successful, or reliable, the system will be.

Paper 56-450, recommended by the AIEE Air Transportation Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE South West District Meeting, Dallas, Tex., April 2-4, 1956. Manuscript submitted December 27, 1955; made available for printing March 7, 1956.

B. J. WILSON is with the Naval Research Laboratory, Washington, D. C.

Alternately, a required or predetermined reliability is often hypothesized. In this sense, the assigned reliability serves as a goal for system design. It represents a level of performance achievement that must be met. In particular, the discussion here is directed along these lines, i.e., of using reliability as a design goal. In addition, the reliability concept formulated here furnishes a logic for analyzing the reliability of each subsystem and equipment of the electric system. Specifically, this problem is approached from the viewpoint that a specified reliability is required of the system. Knowing this, it can be shown how specific equipment is analyzed for individual reliability, and how the integrated assemblage of all equipments is evaluated for over-all system reliability.

Method

The actual process of analyzing electric system reliability is complex. Briefly, it consists first in performing special tests and calculations on individual equipments; i.e., equipment performance is observed for distinct values of applied steady-state voltage and frequency. There are variations in ways of doing this, but the key purpose here is to account for the dispersion or unrepeatability of output performance at repeated but constant electric input conditions. This is conventionally accomplished by calculating from the test results an estimate of the certainty that performance will lie between predetermined limits.

Following these calculations, each of the probable equipment performances is combined or superimposed on a common reference base of voltage or frequency. This superposition is limited to the load equipments with the view of determining the voltage or frequency band that satisfies necessary conditions for reliable operation simultaneously on all loads. Of course, if the presence of one equipment significantly changes the performance of another from that initially measured at given voltages and frequencies, then this superposition is invalid. Such interactions are assumed here to produce only secondary effects in this first approximation of determining system reliability.

Naturally, the limits in voltage and frequency defined by the foregoing procedure must be furnished by generation or other source equipment. But with fixed adjustments such equipment also exhibits a dispersion in performance upon repeated application or use. So the probability of its performing between the required limits must also be evaluated by test and calculation.

It is not sufficient that, by a combined or mockup testing of all system equipments, a single value of voltage or frequency be determined at which the system functions satisfactorily. No such voltage or frequency will ever be maintained perfectly constant. The only other alternative is equivalent to that already employed in the foregoing discussion, where a detailed scheme for analyzing both individual equipments and their later combination is considered.

The method of analyzing the effects either of voltage or frequency is the same as the methods set forth here. So for the sake of simplicity, the procedures to follow will be discussed in terms only of the voltage variable.

For purposes of this discussion, the over-all system reliability will be defined as the product of the independent performance reliabilities of the system equipments. All such independently performing equipments are assumed critical in over-all missile operation. Excessive departure from this independence lessens the efficacy of the definition. Actually, this process of multiplying individual equipment reliabilities ordinarily yields a pessimistic result. That is, the system reliability as a joint probability of equipment reliabilities, inclusive of interaction, generally is greater than that taken as the product of independent reliabilities. Evidence of this rests in the performance of existing complex equipment the actual performance reliability of which is higher than would be predicted from multiplying component part reliabilities. As mentioned previously, only second-order effects are expected from nonindependence, thus permitting reasonably unrestricted use of the reliability definition. So the operational reliabilities p_i of the equipments may be related to over-all system reliability R by $R = p_1\, p_2 \ldots p_a \ldots p_k$.

Since the electric system is related serially to other systems of a missile, such as propulsion and structure, the general approach to reliability in the discussion can be applied conceptually, if not practically, to over-all missile performance reliability.

The Meaning of Failure

A failure in performance results when voltage or frequency variations produce equipment operation outside of acceptable limits. Ordinarily, this type of failure is considered differently than outright cessation of equipment operation. But the fact that it can abort the mission of a missile, as can any other type of failure, ranks it in the same order of importance as more severe equipment failures. These latter often are described as vulnerability failures caused by combat action and degradation failures imposed by shipping and handling.

Analytical Tools

The concept to evaluate electric system reliability is essentially an adaptation of techniques presented in more rigorous discussions on the subject.[1–4] It combines some of the basic statistical ideas of reliability logic and simple superposition of equipment performance and is designed to provide a convenient evaluation for the systems engineer. Statistical evaluations required to qualify performance reliability are held to a minimum, but extension to more generalized conditions can be pursued at the discretion of the engineer with no basic change in method.

Reliability of Individual Load Equipment

The first step in evaluating equipment reliability consists in learning the performance requirements and limits of this equipment. These performance limits define the objectives of the equipment operation and so form the proper basis on which to establish reliability. Any number of influences may cause the limits to be exceeded, thus creating unreliable operation. For example, at a specified input signal and power supply voltage, an amplifier may be expected to produce an output voltage within certain limits. But noise at the input, excessive temperature ambients, or changes in supply voltage may prevent achieving these required limits. In addition, amplifier output voltages may sometimes fall within required limits when the amplifier is operated, sometimes not.

It is possible to draw upon some simple methods of probability and statistics to reduce the aforementioned performance changes to quantitative numbers. By reducing the results of repeated tests on an equipment, its probable performance may be derived and presented in the form of the so-called frequency distribution. For example, the amplifier mentioned previously may be turned on and left on over a specified length of time during which values of output voltage are observed. Either an average value, or values are critical times during operation, are used to record amplifier performance during operation. Generally, the results of all tests, roughly numbering

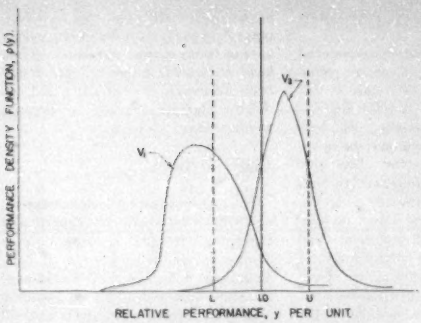


Fig. 1. Probable performance distributions for equipment

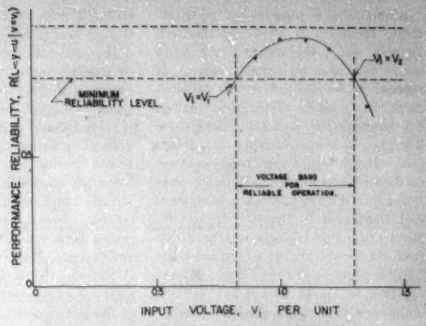


Fig. 2. Equipment reliability characteristics

between 20 to 30, will be scattered about some average or mean value. In a simple equipment, such as an ordinary audio amplifier, the dispersion or scatter of test results will be small; in more complex equipment, such as over-all radar systems, it will be much larger. Ideally, if all conditions of operation exist at their nominal level, this mean value of test performance and the desired nominal performance will be coincident. But this rarely occurs. For convenience, the nominal performance level is assigned the value of 1.0 per unit. Defined in this manner, the performance base is common to all system equipments.

A plot of the frequency distribution which defines the probable performance may be obtained by applying the equipment output test values to equation 1 and by plotting the results against actual performance. The smoothed curve, drawn through discrete points of the discrete frequency distribution approximates the so-called probability density function or continuous frequency distribution $p(y)$.

Fig. 1 illustrates continuous frequency distributions for two values of system voltage V_1 and V_2. For example, had the supply voltage of the amplifier been kept constant at V_1, the V_1 curve of Fig. 1 would have resulted from the tests and calculations just described. The same applies for a different voltage V_2.

The equipment reliability, i.e., the probability that it will perform within its given task limits, is determined by summing up probable performance between these limits. That is, the values expressed by equation 1 are summed between the lower and upper limits of performance L and U which have been assigned according to the performance objective of the equipment. This is shown by equation 2. Most likely this process will entail a graphical integration. Except in rare instances, the distribution of curves illustrated by Fig. 1 is not normal or according to common probability densities. However, curve fitting may prove feasible if automatic computation is subsequently employed.

The reliability so derived will apply only for the one value of voltage to which the curves of Fig. 1 apply. For example, if the curves of Fig. 1 are applied to the example of the amplifier, two different values of reliability will result.

Actually, this fact makes it possible to determine the allowable voltage deviation limits at the input of the equipment. Discrete reliability values may be determined for several voltages V_i. By plotting these values as a function of V_i, the curve shown in Fig. 2 results. By fixing a lower limit of performance reliability, the voltage limits necessary to maintain reliability above this lower limit may be determined. The conditional probability form of equation 3 expresses analytically the manner in which the voltage V_i is handled as a parameter.

Relationship Between Equipment and System Reliability

The sensitivity of over-all system reliability to the voltage or range of voltage becomes apparent when the reliability evaluation of each equipment is completed. If for the moment, the reliability of each equipment is considered to be evaluated at a single voltage V_i, then the system reliability is given by the product of all the equipment reliabilities. This is shown by equation 4. But since each equipment exhibits adequate reliability over a band of voltages, it is necessary first to determine a band common to ay equipments, and one that permits reliabll operation of all equipment. System ree liability will then be derived by equation 4 from R values within this band.

To do this, the results of calculations according to equation 3 for three equipments are superimposed on the common voltage scale in Fig. 3. As indicated there, the common voltage band for reliable operation of all equipments is selected. In this figure, the minimum reliability level is assumed to be the same for all equipments. This is not necessary.

The common voltage band derived from Fig. 3 determines the performance limits for the generation or other source equipment. Thus, generator reliability is evaluated by equation 2, where y takes on voltage values. For the moment, no other parameter is considered. The reliability level for the electric source equipment is shown in Fig. 3.

As observed from Fig. 3, there is a question regarding which equipment reliabilities are to be selected for use in equation 4 for all equipments but the source. If the minimum reliability level is 95 per cent or above, this problem is a minor one. The lowest product obtainable is the safest one to use, which is a so-called pessimistic estimate. A more general expression than equation 3 to include randomness of V_i within the allowable voltage limits is derived in equations 5 through 7.

The equipment reliability evaluations have an added utility. If an equipment does not have a voltage range for reliable operation common to that of other system equipments, it may be considered incompatible for system application and must undergo redesign.

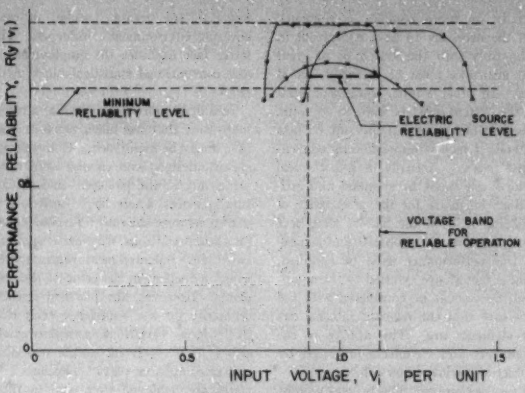


Fig. 3. Superimposed reliability characteristics

Effects of Operational Environments

To an increasing extent among missiles, the effects on internal circuit operation of external environments may have greater influence than normal internal disturbances. This discussion will be expanded to include some of these environmental parameters, specifically temperature and vibration. Time interval of equipment operation may also be considered a parameter. In this last consideration the fact is recognized that equipment may be designed to operate for only short periods of time, for a so-called 1-shot operation. Since such design has little precedent, it must be carefully checked in reliability evaluations. The status of such a practice is highly unsettled and it has advocates and adversaries to its application in design.

Fig. 4 shows the variation in performance frequency distribution with different values of the temperature parameter T_j. Actually, this figure illustrates the classical one-factor-at-a-time variation.

Equation 8 illustrates how temperature parameters may be included in the analytical expression for equipment reliability. This equation is simply an extended form of the original conditional probability form of equation 3. Fortunately, this conditional probability form is well suited to plotting of curves in families as illustrated in Fig. 4.

For all practical purposes, the effect of varying the temperature parameter T_j is to broaden the reliability characteristic R of Fig. 3 into bands. Fig. 5 shows one such band. Here it is shown that at a given voltage level the effect of changing temperature is equivalent to raising or lowering the reliability R. This is also clearly apparent from Fig. 4 when each frequency distribution is summed between fixed performance limits. As observed from Fig. 5, the allowable voltage limits are the narrowest ones formed by the intersection of the reliability characteristic with the lower reliability limit.

The presence of several environments produces R values within or without the band already formed by the temperature parameter. For all intensive purposes, the final equipment characteristic is affected only to the extent further parameter changes broaden it. In the event of too large a characteristic spread or broadening, an equipment no longer qualifies for system use. It performs the task as intended with too little certainty. Such an enlargened band is a signal against further use of the associated equipment. Thus, R bands for all system equipments necessarily are relatively narrow and their superposition continues to be as simple as previously described.

An Alternative Method for Determining Reliability

Alternative methods have been proposed to resolve the effects of environments on equipment reliability.[6-7] In practice, these methods have been confined largely to the elementary component parts. The discussion of one of these methods is included here to orient it in relation to techniques outlined in this paper.

The basis for this method rests on submitting the component part to a severity

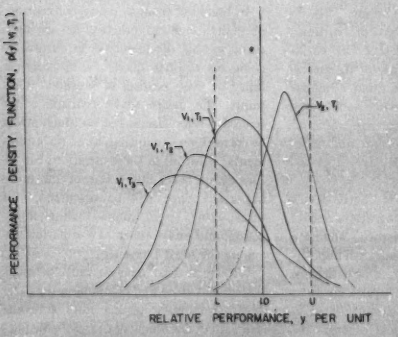


Fig. 4. Performance distributions with variations in a single environmental parameter

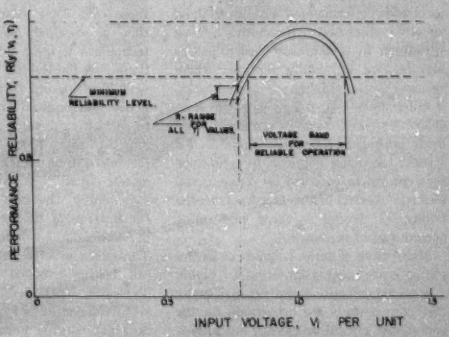


Fig. 5. Effect of voltage band of a single environment variation

of environment several times beyond that expected in actual missile operation.[5,6] In some instances, tests to destruction are actually recommended. This latter criterion is often questionable since an environment may cause a failure to occur, other than an outright destruction, and yet such an environment could create missile unreliability.

In this testing-to-failure method, the reference level for any particular environment is assumed to be that expected from missile operation. The actual environmental or stress level at which failure occurs is noted for each component part. Of course, it is possible to extend such testing to entire equipments. A frequency distribution of the stress levels at which failures occurred is formed from the test points obtained from many component parts. The resultant reliability is considered to be proportional to the value of mean failure level previously established. An excess in mean failure from this reference level of from five to ten standard deviations is considered satisfactory.[7]

Tests based on estimated environmental conditions quite probably could be meaningless. Much uncertainty exists about missile environments both in nature and magnitude. For instance, laboratory vibration tests that sufficiently simulate missile flight conditions have not yet been defined. Exact simulation or sufficient equivalence in duplicating missile conditions may never be practical in the laboratory. However, approximations will evolve to compromise testing adequacy with time and money limitations.

Effect of Production-Run Equipment

Random use of production-run equipment adds performance variations to those already discussed. Estimating these variations in preliminary design phases may be neither possible nor wise where systems perform highly specialized tasks. A tailor-made approach is more direct to achieve specific tasks. Therefore, equipment intended to be highly reliable for system use must conform to conditions that cannot be specified by conventional means. Neglect of this fact has been the failing of several systems performing such duties as fire control.

Microwave systems provide an example of component adaptation to highly specialized circuit functions. Usually their components are adjustable for the commonly critical frequency or impedance factors. Similar adjustments for environments are not easy to provide, especially on more complex equipments. But the necessity to have equipment fit compatibly into the system is so great that insurance must be provided that it does. In other words, equipment built for the system must be built to conform. Specifications must call this out by new methods if such is impossible by conventional means. Actually, a hypothetical performance must be specified and estimations be made for the probability of achieving it. All the factors that bear significantly on the probability of achieving this performance must be specified. If these factors are external to the equipment, they must be considered with the same care that the relevant internal circuit elements are. This applies to environments and to effects produced by production variations as well.

If production variations become a problem, they should be resolved at the manufacturing output point. Referred to as quality control, this process must recognize specifications for reliability that at present have not been formulated adequately to assure the full reliability which equipment should contribute in subsequent missile use. It appears perfectly feasible to formulate suitable specifications. In concept at least, this plan solves the problem of incurring unreliability caused by randomness among production run equipment.

Alternate Analyses for Effects of Parameter Variation and Interaction

An efficient method employed to observe the effects of parameter variation and interaction is referred to as statistical block design. Generally, the results of its application are limited to deciding whether or not variations are greater than those caused by chance alone. To do this, a so-called analysis of variance techniques is used to detect the variations. In the limited sense of its usefulness, block design affords great economy in test data required but, when generalized to furnish equivalent information, economy in using block design and one-factor-at-a-time methods are approximately equal.

Generally, the range of factor levels encountered in missile operation results in well-defined performance changes. These need no sensitive statistical hypothesis testing afforded by block design. Typically, block design is limited to a statistical testing of hypothesis and per se does not provide numerical answers for reliability. Furthermore, the effect of factor variation on performance is nonlinear in the greater number of cases covering environmental influence. This latter fact excludes the employment of any conventional statistical block methods.

In a little more positive vein, a recent innovation that has taken place in block design may be mentioned. Conventional classical designs were created to fill needs principally in the biological and agricultural sciences where large random-type variations were the rule. To compensate for these variations, classical designs require many repeated measurements to effect a reduction in the error of the estimates. However, the physical sciences ordinarily do not experience such relatively large variations or measurement errors, particularly for simple structures such as component parts. Excellent estimates are achieved from one to three measurements. The so-called chain block design, described recently by Youden and Conner,[8] takes advantage of the greater precision among physical and chemical sciences by prescribing one or two replications. The method has been generalized to two dimensions by Mandel.[9] Details of the mechanics for performing analyses associated with block design are provided by Mood and Hoel in references 10 and 11, respectively.

These points are offered to emphasize what type of results may be expected from use of the block design techniques. Restrictions of the application of these techniques have been omitted in many instances where its use has been recommended.[12–13] Aware of these limitations, the system reliability engineer can choose the most effective experimental design and analysis.

Limitations

In breaking from as many conventions of electric system practice as are indicated in this discussion, it is natural to come up against questions of practicability or even realism of method. Certainly, the desire to be realistic is intended. Why else would so much stress be placed on analyzing built-up equipments? Not all of these questions can be considered here, but a few of the more important ones will be discussed. Of these, aging, small sample testing, interaction, and environmental simulation are among the most prominent.

With regard to aging, it must be mentioned that tube and component part use during reliability testing may substantially affect the shape of the probable performance characteristics shown in Fig. 1. In other words, results of the repeated

tests may vary as a function of time lapse throughout the course of the tests, in addition to the natural variability of the equipment.

To analyze this question briefly, two types of equipment may be considered; one type is designed for long-term operation, the other for short-term operation. Most missile equipment is capable of operating for times greatly in excess of expected flight periods. The need for careful preflight checkout supposedly justifies this long-life prerequisite. With this inherent long life, preflight reliability tests should not substantially age component parts, or at least need not do so. In addition, failures in these equipment types diminish exponentially and then taper off in the use period immediately following the first use. Naturally, this presupposes replacement of failed parts. Sometimes a burn-in operation is employed to establish the reduced failure rate prior to application of equipment in the missile. It is during this low failure-rate period that an approximate statistically stationary condition is most apt to be obtained, a condition which is assumed to hold for these reliability tests. Of course, excessive testing may incur damage and cause the failure rate to rise.

For equipment designed to operate solely for the period of flight, the so-called 1-shot operation, the laboratory tests must predict from one set of equipments, the performance in flight of another set. For this situation, it is of paramount importance to duplicate exactly the equipment tested and the equipment flown and to produce equivalent flight conditions. Aging, in the sense of changing characteristics during use, generally has little effect in such instances, even though equipment may be operated once or twice prior to missile installation to determine whether it works.

Statistically speaking, it is true that results from the small sample size employed in the methods discussed here give little assurance of results expected among the entire population of equipments. The emphasis is placed on getting only the prototype system itself to perform reliably in the sense of its dynamic function, and assessing the probability of its doing this. The viewpoint here is that once dynamic performance is achieved to the desired degree of reliability, with the breadboard of prototype system, the results will be used to control the manufacturing of subsequent equipment.

It is true that this viewpoint is a narrow one, limited more or less to a single system with a single purpose. But the intensive or specific task aspect of reliability imposes this condition. In fact, this concept of restrictive use of equipment at first glance is contrary to past practice in aircraft and other electric systems. There the objectives have been to provide equipment usable in the greatest possible number of installations. In missiles the need for generality of application is much less, although the dependability to do a given job must be much greater than for manned aircraft. In the latter, almost numberless redundant functions may be provided by human intelligence to retain satisfactory flight. In missile engineering, the emphasis on equipment is shifted to require almost perfect dependability or, to be more exact, reliability.

Accounting for every variable in missile operation is beyond the present capacity of science and industry. The number of variables are too great and the extent of their interaction is unknown. Although interaction is a deterrent to predicting system performance, it should not prevent this prediction. Since interactions may be detected, at least their effects may be noted before missile operational use and remedies taken to correct effects harmful to missile performance. For example, interference signals coupled through the electric system may enter the input of the amplifier previously discussed to create an improper output or even produce oscillations. Such anomalies may be corrected by filtering, shielding, or other decoupling measures.

Interaction may be less apparent than in the previous example and yet seriously affect system reliability. For instance, it has been assumed that all equipments perform independently. But varying currents through the line impedances between equipments may create coupling voltages. Thus, independence of performance is diminished although it may not seriously affect system reliability. Coupling such as this creates a discrepancy between the reliability calculated as a product of independent equipment reliabilities and the actual or measured reliability of performance. It is possible to treat such interactions both analytically and experimentally but doing so is beyond the scope of this discussion.

It has been mentioned that duplicating missile operational environments in the laboratory is exceedingly difficult, and in some cases, viz., vibration, nearly impossible. In particular, it is felt that these environments should be applied simultaneously, principally to assure that effects of interaction among environments are not overlooked. Fortunately, the intimate effect of environments on circuit operation is receiving increasing attention and, although today's reliability studies account for environmental effects rather inadequately, results of such studies are accordingly considered conditional until better simulation techniques are developed. In the meanwhile, failures reported from actual flight, when learned, can be correlated to laboratory test failures. From an engineer's viewpoint, however, this procedure is unsatisfactory, and a method of predicting operational reliability in the laboratory is more economical in time and expense.

Conclusions

The approach on system reliability outlined here attempts to serve two masters. One would have the methods indicated tied closely to realizable and economical designs and tests; the other would have them formulated consistent with the over-all logic of reliability. To satisfy the former, the entire discussion is predicated on tests from one each of the existing system equipments, but only with the shortcomings of such a procedure in full view. To appease the latter, probability principles are applied where possible.

In general, the equations perform as guides rather than as plug-in formulas. That is, the main body of calculation consists in reducing test data to discrete values, drawing smooth curves through them, and proceeding on a graphical basis.

To date, the scope of the reliability problem has been largely confined to consideration of individual equipment. This has been complex enough. Expanding it into the realm of entire systems does not present an inviting prospect. In addition, the solutions must have great utility and even be useful in integrating complete systems from existing equipments.

The methods developed in the foregoing have been designed to serve the following functions:

1. To define the voltage tolerances that must be maintained when all equipments work together. Akin to this, they define the boundary conditions of voltage common to both electric sources and loads. An analogous process may be applied to frequency where explicit interpretation in these terms is preferred.

2. To indicate which equipments are least reliable and/or compatible and should undergo redesign.

3. To permit evaluation of numerical equipment and system reliabilities inclusive of the source.

4. To allow for determination of environmental influences.

5. To provide a basis for analysis unlimited by complexity or performance objectives.

6. To lead to eventual correlation between component part and over-all equipment reliabilities as the latter are produced in increasing numbers.

7. To give quality control, applied during manufacture, those equipment performance limits that are necessary to check whether or not production-run equipments perform within limits specified by reliability requirements.

8. To permeate the activities of purchasing, procurement, and manufacture by accurate and fundamental specification of performance requirements that apply in these areas.

The first function is vital. It involves the dependence in performance of equipment developed in one geographical location upon that developed independently in another and assists in predicting how compatibly they will work together in a common system. As intimately related as missile subsystems and component would seem to be, they are not always so treated. The second function of selecting as to which equipments to redesign resolves an otherwise critical problem.

The evaluation of numerical reliabilities is the goal of any effective reliability effort. The method outlined here recognizes and avoids the formidability of direct analytic attack in the sense of so-called exact mathematical models. For such would be an almost impossible task even without the necessity of including environmental effects.

The fact remains that this analysis does not reach all areas necessary for reliability improvement. The basic building blocks for the equipment are the component parts, the resistors, relays, vacuum tubes, transformers, and solid-state electronic devices. In their improvement lies the secret of achieving one of the fundamental steps of reliability. Another basic step is proper application of these parts. There is great need to speed up the correlation between component part and equipment reliabilities, but doing this is beyond the province of this discussion. Figuratively, the input-output viewpoint of an entire equipment is macroscopic in nature; by contrast, the component part viewpoint is microscopic. But the effect of the latter on the former is inseparable and should be monitored as more and more over-all equipments are produced.

The last function in the foregoing list touches on a controversial area. Usually, the areas of purchasing and procurement lie outside the reliability engineer's authority. But they do not lie outside his responsibility. For example, as a tool, reliability has a bilateral action. The engineer must usually rely on parts made elsewhere to synthesize his complex equipment. But the reliability of the outgoing equipment is in jeopardy unless the incoming parts are reliable. Inappreciation of this fact on the part of the procurement agent exposes the engineer's efforts to failure. So the true extent of the engineer's action is clear. He must specify his needs unambiguously. Hence, the inclusion of reliability aspects in the general procurement specification is requisite. Their enforcement is the engineer's professional responsibility.

Appendix

Nomenclature

y = relative performance, per unit
y_i = discrete value of relative performance
$\Delta_i y$ = incremental interval i of relative performance
$n_{i_y \Delta}$ = number of discrete performance values in $\Delta_i y$
n = total number of performance test samples
s = total number of system equipments
$p_i(y)$ = relative performance frequency for the ith interval
$p(y)$ = probability density function or probable performance, the continuous frequency distribution derived from the values
L = lower limit of probable performance $p(y)$ at $y = L$
U = upper limit of probable performance, $p(y)$ at $y = U$
R = performance reliability
V_i = the fixed value of steady-state applied or input voltage
T_j = parameter designating temperature level

Calculations

The frequency distribution for discrete variable values may be expressed as

$$p_i(y) = n_{\Delta_i y} \tag{1}$$

An approximately probability density function $p(y)$ is formed when a smooth curve is drawn through the point $p_i(y)$.

Equation 2 describes the net probability of occurrence of y over a performance interval $L < y < U$

$$R(L<y<U) = \int_L^U p(y)\,dy \tag{2}$$

By the introduction of the voltage parameter V_i, equation 2 converts to the conditional form

$$R(L<y<U|V_i) = \int_L^U p(y|V_i)\,dy;\ i = 1,2,3 \ldots m \tag{3}$$

The results of a series of calculations by equation 3 are generally regarded as discrete values.

If a system of equipments are serially related, i.e., the system is critically dependent on all of them, and each performs independently of the others, the over-all system reliability may be written

$$R_{system} = R_1 R_2 R_{secure} \ldots R_s \tag{4}$$

Equation 3 may be generalized to include the effects of a random variation in V_i. This is expressed as

$$R(L<y<U, V_L<V_i<V_U) = \int_{V_L}^{V_U}\int_L^U \times p(y, V_i)\,dy\,dV_i \tag{5}$$

Since the joint distribution

$$p(y, V_i) = [p(y|V_i)][p(V_i)] \tag{6}$$

where p (V_i) is the marginal distribution $\int_U^U p(y, V_i)dy$

$$R' = \int_L^U \int_{V_L}^{V_U} [p(y|V_i)][p(V_i)\,dV_i]dy \tag{7}$$

where R' is the equipment reliability modified to account for randomness in V_i. Since $R = \int_L^U p(y|V_i)dy$ already has been evaluated in equation 7, R is weighted by the marginal density $p(V_i)$.

Equation 3 may be expanded to include other parameters. For example, if T_j is introduced as a temperature parameter, equation 3 becomes

$$R(L<y<U|V_i, T_j) = \int_L^U p(y|V_i,\ T_j)dy \tag{8}$$

References

1. A Survey of the Current Status of the Electronic Reliability Problem, R. R. Carhart. *Report RM-1131*, Rand Corporation, Santa Monica, Calif., Aug. 1953.

2. Methods for Calculating Reliability of Pilotless Aircraft, C. Gates. *Proceedings*, Airborne Electronics Conference, Dayton, Ohio, 1952, pp. 9–10.

3. On the Reliability of Networks, G. H. Weiss, M. H. Kleinerman. *Proceedings*, National Electronics Conference, Chicago, Ill., vol. 10, 1954, pp. 128–35.

4. Designing Reliability into Electronic Circuits, A. H. Bonner, B. Meredith. *Ibid.*, vol. 10, 1954, pp. 137–45.

5. A Study of the Methods for Achieving Reliability of Guided Missiles, R. Lusser. *NAMTC Technical Report 75*, Naval Air Missile Test Center, Washington, D. C., July 19, 1950.

6. Reliability Specifications for Guided Missiles, A Proposal, R. Lusser. Redstone Arsenal, Huntsville, Ala., Sept. 1955.

7. Reliability Improvement—A Plan to Achieve and Measure Reliability of Aircraft Equipment, J. E. Luckman. *AIEE Transactions*, vol. 73, pt. II, 1954 (Jan. 1955 section), pp. 378–83.

8. The Chain Block Design, W. J. Youden, W. S. Connor. *Biometrics*, New Haven, Conn., vol. 9, June 1953, pp. 127–40.

9. Chain Block Design with Two-Way Elimination of Heterogeneity, J. Mandel. *Ibid.*, vol. 10, June 1954, pp. 250–72.

10. Introduction to the Theory of Statistics (book), A. F. Mood. "Experimental Designs and the Analysis of Variance." McGraw-Hill Book Company, Inc., New York, N. Y., 1950, pp. 316–58.

11. Introduction to Mathematical Statistics (book), P. G. Hoel. "Application of the F-Distribution." John Wiley & Sons, Inc., New York, N. Y., 1947, pp. 152–61.

12. The Necessity of Statistical Experimental Design in Testing for Component Reliability, J. L. Blair. *Proceedings*, Electronic Components Symposium, Pasadena, Calif., 1953, pp. 20–1 to 20–7.

13. Some Experience with the Design of Experiments, J. R. Duffett. *Industrial Quality Control*, Milwaukee, Wis., vol. XI, no. 3, Nov. 1954, pp. 36–40.

Discussion

Cyrus J. Creveling (Naval Research Laboratory, Washington, D. C.): The term reliability throughout most of this paper is used in a slightly different manner than in its most generally accepted definition. According to the context and as defined by equation 2, performance reliability is the probability that performance will be found to be within certain limits when the equipment is subjected to certain variations in environmental parameters. This performance reliability then may be computed from certain tests on the equipment before use. This appears to be preliminary to the conventional defining and measuring of reliability as a function of time, as set forth in reference 1 of the paper.

Most of the tests to determine the probability of correct operation, as described by the author, such as varying voltage, temperature, etc., are necessarily applied in the design stage to insure compatibility and, ideally, should allow introduction of sufficient tolerances so that these factors can be virtually ignored thereafter.

In the design of conventional equipment, in which long life is desired, the aforementioned tests are made as a matter of course and are taken for granted when attempting to improve reliability by other means.

On the other hand, I feel that missile designers, since they are used to dealing with short periods of time, may feel that time itself can be ignored, and the quasi-stationary state consequently receives most of the attention.

Still, discussions on missiles refer to failures which are attributed to random causes. These failures affect reliability in the time domain. This means that an equipment, however well engineered, cannot be completely relied on just because it has just been tested and found to be working, no matter how thorough these tests may have been. Failures will still occur and their probability can only be predicted on the basis of life tests. It is true that the 1-shot type of operation may practically eliminate consideration of the wearout characteristics of most electronic equipment, especially those which result in a rectangular time function of probability. This leaves the random effects, however, and those are usually the predominant ones in failure of electronic equipment. This being the case, the reliability of missile or other 1-shot applications, as it is a function of time, must still be considered in its over-all reliability. After testing techniques and equipment design are perfected, these random effects will still affect life and reliability, and other principles such as life testing and redundancy may still be necessary to maintain an initial high probability of success.

I believe that the several approaches to increasing reliability, i.e., quality control of components, system testing for compatibility, type testing for life, life testing under use or simulated use, and the use of redundant parts and systems, are all effective tools for improving reliability, and that all are necessary if the maximum reliability is to be achieved. It seems clear that any reliability program which ignores life testing with successive modification will fail to achieve maximum results. The quality control programs are achieving steady improvement in reliability, but this is a long-time proposition, and it may be a long time in producing the orders of magnitude of improvement which are needed in some applications.

A seasoned design is arrived at in conventional equipment by building developmental equipment which can be modified and improved as the result of laboratory tests and as the result of experience obtained in use. These are the so-called feedback loops. The 1-shot type of operation of most missiles has discouraged some missile designers from thinking that these feedback loops can be closed. I believe, however, that a program of life testing under simulated operating conditions may perform the functions of synthetically accomplishing this. Instead of accepting missile components from the factory and testing them and then saving them for use, they could be put in simulated use until actually used. The failures caused by simulated use would then indicate design weaknesses which could be corrected in manufacture. This would result in the evolution of a proved design in much the same manner as the process used in the development of conventional equipment, and would provide failure data for an assessment of time reliability.

It has puzzled me for some time that in the field of reliability there are a number of theories as to how reliability can best be achieved. Furthermore, there exists something of a controversy as to the merits of some of the means, particularly in regards to redundancy. Although an advocate of redundancy, I do not doubt the need of persuing reliability along lines of component improvement, quality control, life tests, circuit development and refinement, and along the lines of systematic tests as outlined by the author. My belief is that, necessary and useful though these approaches be, they are incapable of supplying the required reliability and will still require the use of redundancy. I might add that there has been expressed the opinion that redundancy is not only unnecessary, but actually is a fallacious concept, the adoption of which will decrease reliability. These opinions may have their origins in the belief that in missile design the random effects of time are negligible. If this is true, redundancy has little to offer in increasing reliability. I believe however that this discussion may serve to help reconcile these diversities of opinion by showing that time, however short, is the essence and that redundancy is one means of combatting its effects. There may also be disagreements as to whether the premises upon which redundancy is based may be realized in practice, but I believe that techniques known today need merely be applied.

B. J. Wilson: It is advisable to distinguish between randomness in performance and randomness of failures. The former must be accounted for to have a true measure of quantitative performance; the latter is not necessarily related to performance at all except as a special case. Among missiles, failure is a performance that cannot be tolerated, and the singular knowledge needed is how certain we are that failures won't occur during flight. But this is only one part of the knowledge to be gained by performance reliability. Generally, the tests supplied during design are not sufficient to correlate quantitative performances that occur because of the randomness effects produced by the combination of many parts which exhibit hysteresis and wide performance spread.

Regarding reliability as a function of time, the paper under discussion does imply that the failure rate, i.e., the proportion of random failures per unit time not the proportion of failures per thousand component parts, becomes negligible during shelf life subsequent to test and prior to use in flight, and that random failure rate as well as wearout remain somewhat stationary during reliability evaluation test failure. Neither of these ideal conditions prevail. For instance, consider the deterioration of electrolytic capacitors during storage. On certain limited occasions, it may prove desirable to sample periodically the performance of stored equipment and fold in the resultant distribution with the random performance distribution. Nevertheless, test samples as prescribed in the paper are held to have constructive value.

No doubt continuous use with continuous monitoring of performance of every part or equipment from the time of manufacture to the time of flight would result in a better reliability prediction. But many factors may make this infeasible; for instance, vibration-induced fatigue or temperature-induced deterioration of insulation may occur. Much engineering judgment must go into the design compromises necessary to make equipment economical, efficient, and also reasonably environmental free.

The application of redundancy principles is almost essential to missible reliability. It is doubtful if actual required reliabilities for links in a chainlike system will ever be achieved in view of the ever changing needs. A recent paper[1] on this subject is commendable in view of its clarity and indicates the generality of use for redundancy that is made apparent by Mr. Creveling's[1] constructive applications.

Reference

1. Increasing the Reliability of Electronic Equipment by the Use of Redundant Circuits, C. J. Creveling. *Proceedings*, Institute of Radio Engineers, vol. 44, Apr. 1956, pp. 509–15.

Recent Trends in Aircraft Ground Power Electric Equipment

H. S. SECHRIST
ASSOCIATE MEMBER AIEE

M. ROSENBLATT
ASSOCIATE MEMBER AIEE

C. M. BRYANT
NONMEMBER AIEE

Purpose of Ground Power

AIRCRAFT ground power equipment serves two primary requirements. It makes possible the starting of the airplane's engines, and it provides electric power to the airplane's utilization equipment, thus conserving the plane's own electric system for use during actual flight only. This paper is concerned with the electric equipment used in engine generator sets which provide the source of aircraft ground power.

Nearly all military aircraft and most commercial aircraft depend on an external source of electric power for starting since the weight penalty incurred in carrying batteries of sufficient capacity for cranking engines is too great to be tolerated. As an example of the power required, a jet fighter such as an *F-86* requires 1,000 amperes at 28 volts direct current for starting the engine in the proper period of time. A reciprocating aircraft such as a *DC-6B* requires 300 to 400 amperes for starting as well as 400 to 500 amperes at 28 volts direct current for supplying utilization equipment.

In military applications, the ground power equipment usually operates more hours in servicing the utilization equipment than does the aircraft electric generating system. The newer and larger military aircraft with their complex armament, bombing, and navigation systems require more hours' servicing on the ground for each hour of flight. Military aircraft standing on the alert may require their electric systems to be energized continuously by ground power equipment.

Aircraft Electric Power Requirements

In general, aircraft electric power requirements can be broken down into three categories according to aircraft:

1. Aircraft requiring only d-c power.
2. Aircraft requiring d-c and 400-cycle a-c power.
3. Aircraft requiring only 400-cycle a-c power.

Historically, the trend of power requirements appears to be in that order. Commercial aircraft fall into the first category; a number of the military aircraft in the fighter class also require only d-c power. Some of the latest military aircraft require only a-c power; this is made possible by starting the engines with air or combustion starters.

Types of Ground Power Equipment

Probably the first aircraft ground power supply was the battery cart for supplying d-c electric power. In the past, ground power for military aircraft has been primarily d-c with small amounts of a-c power. The commercial airlines have required d-c ground power only. There have been a number of ground power units in use by the Air Force over the past several years with the output power available shown in Table I.

Several years ago, the Navy standardized a general specification for engine generator sets, *MIL-P-6905A (AER)*. This specification covers both a-c and d-c systems. The first ground power supply built to meet these requirements was the *NC-5*, shown in Fig. 1. These units supplied 15 kw of direct current with a rotating generator, and 6 kw of direct current with a transformer rectifier unit. The a-c generator was rated 30 kva.

Importance of Reliability and Life

The need for equipment with long life becomes apparent in consideration of the long periods over which the ground power equipment must be operated. If operated eight hours per day, five days a week, in six months time the equipment will have encountered 1,000 hours of operation, so that generally a minimum of 1,000 hours between overhauls is specified, and usually 5,000 hours is required.

Reliability is very important with many aircraft as any interruption in power may require beginning over again a lengthy calibration procedure requiring several hours. Some equipment, particularly that utilizing gyromotors, may require a time interval before reapplication of power in order that the gyro may decelerate. Power interruption could be very critical on aircraft standing on alert because of the waiting period required after interruption of power and warm-up period required following reapplication of power.

The control devices, regulating systems, and generating equipment must fulfill both the reliability and life requirements. However, they must be small enough and light enough to permit mounting in a minimum of space and without undue weight penalty. Air transportability is a requirement for most engine generator sets, and heavy, bulky equipment imposes a penalty even when transported in an aircraft such as a *C-119*.

To meet weight, space, and environmental requirements, it is usually necessary to use conservatively rated aircraft-type control devices. For example, a National Electrical Manufacturers Association size-4 contactor is generally too bulky and heavy to mount in a small trailer or self-propelled unit. Furthermore, standard commercial control devices give trouble when operating under the severe environmental conditions found in military applications.

The reliability requirements of ground power regulating systems are usually being met by employing the static magnetic-amplifier type of device which also has long life. The life and performance of the electromechanical type of aircraft regulators are usually not satisfactory for ground power applications. Electronic regulating systems are usually not capable of withstanding the shock and vibration normally encountered with self-propelled engine-driven equipment.

The generating equipment for ground power must also be reliable and light in weight, as compared to the industrial types. In the past, light weight could be achieved only by the use of aircraft generators operating at speeds up to 6,000 rpm. Recently, a new semi-industrial type of generator has been designed which operates at lower speeds, such as 3,428 rpm, and which utilizes aluminum frame construction and solid shafts. Shown in Fig. 2(A) is a 60-kva

Paper 56-445, recommended by the AIEE Air Transportation Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE South West District Meeting, Dallas, Tex., April 2–4, 1956. Manuscript submitted September 19, 1955; made available for printing March 7, 1956.

H. S. SECHRIST, M. ROSENBLATT, and C. M. BRYANT are with the General Electric Company, Waynesboro, Va.

Table I. Air Force Ground Power Units

Designation	D-C Power	A-C Power	Engine	Type
A1A	28-volt, 9.4 kw, 11 kw overload		air-cooled gaso-line	4-wheel trailer
C-26	28-volt, 45 kw, 60 kw overload	115-volt 1-phase, 380-980 cps, 15 kw, 20 kva overload	air-cooled gaso-line	3-wheel self-propelled
B10B	28-volt, 7.5 kw	115/200-volt 3-phase, 400 cps	air-cooled gaso-line	3-wheel trailer
	110-volt, 3 kw (TR)*	40 kva, 0.75 PF		
A-4	28-volt, 30 kw, 120-volt, 20 kw		air-cooled gaso-line) *4 units	3-wheel self-propelled
B-47 ground support system	28-volt, 45 kw *4 units	115-volt 1-phase, 400 cps, 20 kva; 115-volt 3-phase, 400 cps, 3 kw; 440-volt 3-phase, 60 cycle, 150 kw — 4 units	Diesel	9-trailer units
MA-1	28-volt, 45 kw, 67.5 kw overload	115/200-volt, 3-phase 400 cps, 30 kva, 0.75 pf	water-cooled gasoline	4-wheel self-propelled driver ride
MD-3	28-volt, 45 kw, 60 kw overload	115/200-volt, 3-phase 400 cps, 60 kva, 0.75 pf	air-cooled gaso-line	3-wheel self-propelled

*Transformer rectifier unit.

Fig. 1. Two views of Navy type, NC-3, ground power self-propelled vehicles

400-cycle 3,428-rpm a-c generator of this type which weighs 450 pounds. In comparison, Fig. 2(B) shows an older design, rated 30 kva, 400 cycles, 3,428 rpm, and weighing 525 pounds, utilizing cast-iron frame construction.

Aircraft alternators are being built today which have longer life than those used previously. Fig. 3 shows a self-ventilated machine rated 60 kva, 400 cycles, 6,000 rpm, in use on ground power equipment at present. The life of bearings constitutes one of the main problems in obtaining maximum life from an alternator. In sealed bearings of the type used on ground power equipment, bearing life is directly dependent on lubricant life. Tests show that the lubricant life decreases approximately 1.5 times for each 10 C (degrees centigrade) increase in bearing temperature. Because aircraft alternators are usually operated with high temperature rises which directly affect bearing temperatures, the expected bearing life of the semi-industrial type of alternators is several times longer than that of aircraft alternators. The aircraft type of alternators utilized on ground power are usually designed to limit the temperature rise through the generator, to meet the life and reliability requirements.

Environmental Conditions

In general, ground power electric equipment must be designed for operation in ambient temperatures of minus 55 to plus 55 C, at altitudes up to 5,000 feet and in relative humidity conditions up to 100% (per cent) where condensation takes place on the equipment. It must be operable after being subjected to sand and dust conditions; salt spray; fungus growth, such as is encountered in tropical climates; and after being subjected to vibration and shock encountered while being transported over all types of roads and while being air transported.

A-C Electric Systems

A-C Load Characteristics

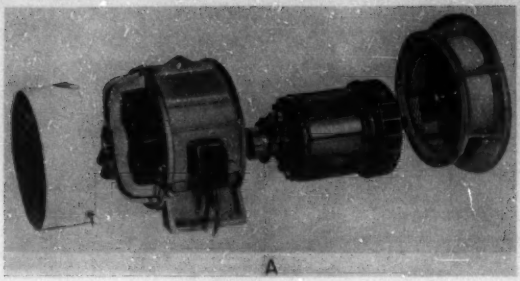
A-c ground power equipment is used to supply electric energy to the aircraft pumps, compressor motors, air-conditioning systems, de-icing equipment, and electronic equipment of all kinds. The a-c ground power equipment is sometimes required to serve as a reference in the calibration of precise electronic equipment. The a-c ground equipment may also be required to operate the planes' transformer rectifier units and the planes' inverters.

Some airplanes may require very small amounts of alternating current (less than 1 kw) and some require as much as 60 kw.

The frequency of the a-c ground power systems is invariably 400 cycles, which is consistent with the plane's electric system. The ground power a-c generators are usually rated 200 volts line to line, 3-phase, 4-wire, which enables the airplanes' single-phase 115-volt loads to be connected line to neutral, to the ground power vehicle.

General Performance Requirements for A-C Systems

The performance requirements for voltage regulation, recovery time, voltage modulation, and wave form on a-c ground power systems for military aircraft have become increasingly rigid over the past year. One-per-cent voltage regulation, 0.15-second recovery time, 1% voltage modulation, and 2% rms total harmonic content have recently become accepted ground power requirements, and are usually more stringent than those specified for the aircraft. The question is frequently asked as to why the performance requirements for the ground power equipment are more rigid than those for the aircraft's electric system. There are several logical answers to this question. It is desirable to calibrate the very complicated electronic equipment on today's military aircraft in the center of the band of voltage variation which may be encountered while the equipment is operating with the aircraft electric system. By making the limits for the ground power equipment narrower than those for the aircraft electric system, it is possible more nearly to approach the median point of voltage variation encountered during actual operation. Another reason for the rigid voltage regulation limits and rapid recovery time requirements is to minimize disturbances to the electronic systems during calibration so that erroneous adjustments will not be made.



A

B

Fig. 2. Semi-industrial 400-cycle 3,428-rpm a-c generators

A—Sixty kva; aluminum frame construction. Approximate dimensions: frame diameter, 19 inches; length (face of flange to end cover), 23 inches; weight, 450 pounds

B—Thirty kva; cast-iron frame construction. Approximate dimensions: frame diameter, 15$^{3}/_{8}$ inches; length (face of flange to end cover), 31$^{3}/_{4}$ inches; weight, 525 pounds

It has been found that poor wave form introduces extraneous disturbances into the electronic equipment, which makes calibration difficult. Thus, the total harmonic content of less than 2% by the rms method is required.

Consideration must be given to maintaining voltage regulation over wide ambient temperature ranges. Operating conditions may be such that the ground power equipment is adjusted inside a hangar and the voltage properly set. The unit may be moved outside where the ambient temperature may be 50 to 75 C different. This operation might well be encountered in polar and tropical regions and because of the type of clothing worn by operating personnel, any adjustment outside the hangar would be difficult. Thus, there is a definite need for maintaining the voltage regulation limits over the ambient temperature range.

A-C System for NC-5 Ground Power Unit

The a-c system used in the *NC-5* ground power equipment consists of a 30-kva 400-cycle 200/115-volt 4-wire semi-industrial a-c generator with an overhung amplidyne exciter, shown in Fig. 2(B). The rotating amplidyne exciter employed is a series-type exciter which will build itself up to ceiling voltage when no current is supplied to its control fields. This type of exciter was chosen because of its inherent high gain (the power required to control it is very small) and because its residual voltage will cause the system to start up without utilizing an external source of energy. Since the amplidyne exciter will automatically build itself up to a high output voltage, it becomes the function of the voltage regulator to "buck" down the amplidyne voltage to the required value so as to maintain the alternator output voltage at the desired value.

A schematic sketch of the system including the voltage regulator is shown in Fig. 4. The input rectifiers *CR2*, *CR3*, and *CR4*, of the voltage regulator supply a direct voltage proportional to the average of the 3-phase line-to-line a-c generator voltages. This direct voltage furnishes a supply for the constant-voltage cold-cathode gas-filled reference tube. A portion of this direct voltage (proportional to the alternating voltage), taken from the voltage divider consisting of *R10*, *R9*, and *R5*, is compared with the constant-voltage reference, and the difference between these two voltages causes a current to flow through *R3* and the control winding of the magnetic amplifier. Should the a-c generator voltage be higher than normal, the current will flow through the magnetic-amplifier control winding in a direction which will increase the output of the magnetic amplifier, causing increased current to flow through the buck *F2-F1* amplidyne field so as to lower the amplidyne output voltage and thus bring the alternating voltage back to normal. The converse takes place for a low a-c generator voltage. A bias field is provided on the amplidyne *F3-F4* to help restore the amplidyne output voltage after the voltage regulator has forced the amplidyne output voltage to zero or even negative, during the process of achieving rapid system response immediately following load removal on the a-c generator. The rate of change of amplidyne armature voltage and the rate of change of magnetic-amplifier output voltage are fed back into the magnetic-amplifier control winding in a degenerative manner, to achieve system stability. The forward voltage drop across the input selenium rectifiers *CR2*, *CR3*, and *CR4* will change nonlinearly with extensive changes in temperature and will introduce an appreciable error into the voltage regulator. The error is compensated by the presence of a similar nonlinear device (selenium rectifier *CR5*) in the direct voltage divider.

The characteristics of the amplidyne exciter are such that as its armature voltage is increased, the magnetic amplifier of the voltage regulator is required to increase its output; therefore, as load is added to the a-c generator, its output voltage will increase. This characteristic tends to compensate for the voltage drop in the 30-foot output cable connected to the ground power unit which helps maintain the alternating voltage at the airplane more nearly constant.

This system maintains the a-c generator voltage within ±2.5 volts line to neutral of the rated value while the system is subjected to load changes between no load and full load, ±5% frequency

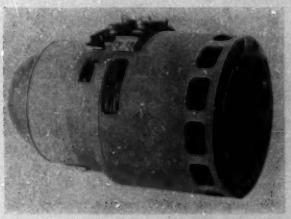


Fig. 3. Aircraft-type 60-kva 400-cycle 6,000-rpm a-c generator, self-ventilated

Approximate dimensions: frame diameter, 11 inches; length (face of flange to end cover), 17$^{11}/_{32}$ inches; weight, 123 pounds

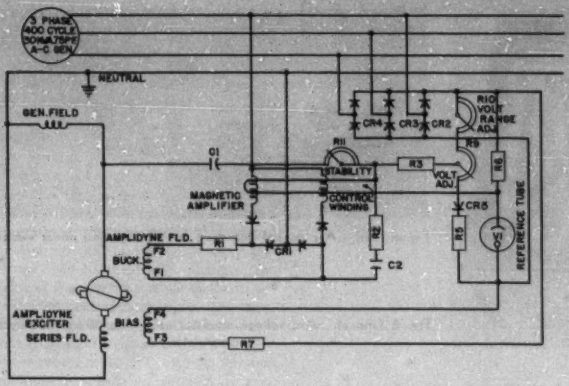


Fig. 4. Circuits of alternating voltage regulator used on NC-5 ground power unit

changes, and all environmental changes. For environmental changes normally encountered, the voltage at the end of the cable is usually held within 1% of the rated value. The system voltage will recover to within 5% of the final value after a 30-kva 0.75 power factor load is suddenly applied or removed in 0.2 second or less. Fig. 5 is a recording showing application and removal of rated load on this system.

In addition to the voltage regulator, the a-c control box for the *NC-5* ground power equipment contains overload protection and phase rotation protection. The equipment also contains a voltmeter for reading all three line-to-neutral voltages and an ammeter for reading all three line currents. Convenience outlets are furnished to supply power for portable a-c equipment. Also, a line contactor is furnished to energize or de-energize the cable to the airplane while the a-c electrical equipment is in operation. The ground power a-c equipment is radio interference suppressed to meet the requirements of *MIL-I-6181*. Fig. 6 shows the voltage regulator and control equipment for the a-c generator used on the *NC-5* ground power vehicle.

A-C System for MD-3 Ground Power Unit

The aircraft type of a-c generator used on this ground power unit is rated 60 kva, 0.75 power factor, 400 cycles, 200/115 volts, 6,000 rpm, and is shown in Fig. 3. This generator is equipped with an overhung conventional exciter. Both the generator and exciter are self-cooled by a fan which is mounted on the armature shaft.

Fig. 7 shows a simplified schematic sketch of the static voltage regulator used to control this 60-kva generator. The 3-phase input rectifiers and the comparison circuit are very similar to the system previously discussed for the *NC-5* equipment. The comparison circuit in this case controls a first-stage magnetic amplifier connected in push-pull. The output of the push-pull first-stage magnetic amplifier controls a single-ended second-stage magnetic amplifier, which in turn supplies power to the field of the rotating exciter. The first-stage magnetic amplifier is connected push-pull in order to minimize the affects of frequency changes and temperature changes on its performance. Two stages of magnetic amplifier were employed because the rotating exciter requires more power to control it than the amplidyne used in the previously discussed system. Since the voltage regulator in this case supplies positive excitation to the rotating exciter, the voltage error in the system will be such that, as load is added to the generator, its voltage will drop slightly. To achieve close regulation limits at the end of the 30-foot output cable, it is necessary to add a line drop compensator to this system. The line drop compensator shown schematically in Fig. 7 can be adjusted so that its constants match those of the cable and thus it will introduce a signal into the input of the voltage regulator which causes the generator voltage to rise the same amount as the voltage drops in the cable.

Because the rotating exciter is a conventional shunt-type d-c machine, it is necessary to employ a relay *ICR* to self-excite the exciter until the a-c generator voltage builds up to a high enough value to energize the relay (about 80% to 90% of rated). The rate of change of exciter armature voltage

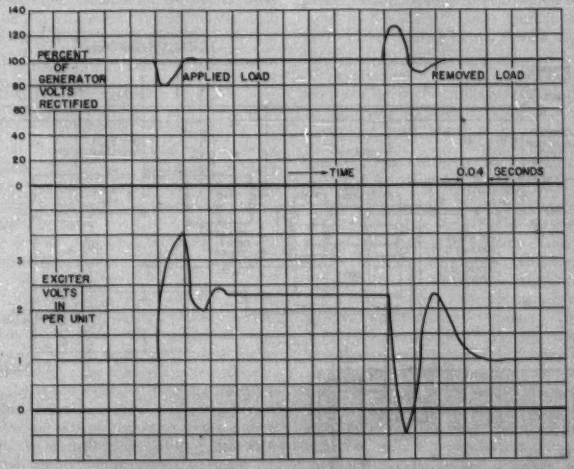


Fig. 5. Recording of rectified a-c generator voltage and a-c generator field voltage during application and removal of rated load on NC-5 a-c system

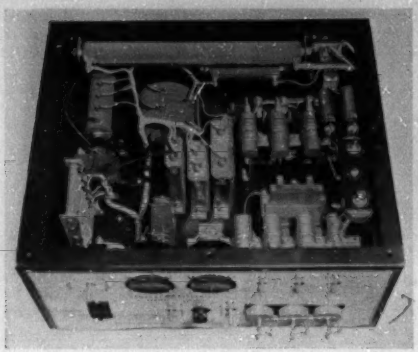


Fig. 6 (left). A-c control box used on NC-5 ground power vehicle

Fig. 8 (above). A-c voltage regulator used with 60-kva 400-cycle a-c generator

and exciter field voltage is used for stabilizing signals.

This system maintains the output voltage to within 1% of the no-load value at the end of the cable during no-load to full-load changes. It restores output voltage to within 1% of the final value in 0.15 second or less. Fig. 8 shows the static alternating voltage regulator used in the *MD-3* ground power vehicle.

Recently Developed A-C System with Static Excitation

A recently developed a-c system with static excitation is used to control the output of the semi-industrial type of generator shown in Fig. 2(A). Fig. 9 shows the circuit used in the static excitation system. Usually a 3-phase potential transformer *1T* is required to match the a-c generator output voltage to the generator's d-c field voltage. In series with the secondary of this transformer are three linear reactors shown as *1X*, *2X*, and *3X*. A 3-phase saturating current transformer *1CT* is employed as the controllable element in the excitation system. A 3-phase full-wave selenium rectifier converts the a-c output of the static excitation system into d-c field power for the generator.

A single-phase equivalent circuit for this type of excitation system is also shown in Fig. 9. In this equivalent circuit, the secondary voltage of the potential transformer is represented by E. X_1 is the equivalent single-phase reactance of the three linear reactors. X_m is the magnetizing impedance of the secondary of the saturable current transformer. The magnitude of this magnetizing impedance is usually comparable to the resistance of the generator field. R in the equivalent circuit represents the resistance of the generator field in terms of its equivalent a-c value. I is a current proportional to the load current of the generator times the turns ratio of the saturable current transformer.

With no load on the a-c generator, I is zero and E causes a current to flow through reactor X_1. This current then divides, and part of it goes through the magnetizing impedance of the saturable current transformer (X_m) and part through R. The component of this current which flows through R, shall be called I_e, this component of current being proportional to generator voltage. When a load current flows in the generator output, a current I is introduced into the equivalent circuit. This current divides, part of it flows through X_1, part through X_m, and part of it through R. The component of this current which flows through R, shall be called I_i, since it is proportional to generator load current.

With no load on the generator, the generator terminal voltage E gives the required generator field current I_f. As load is added to the generator, the current I produces more field current by the addition of the current I_i. I_i adds vectorially to I_e so as to give approximately the correct value of generator excitation. This vectorial addition is at right angles for unity power factor generator loads, and the two currents I_e and I_i add directly for zero power factor lagging generator loads. Thus, the system is designed so that the field current component I_i is

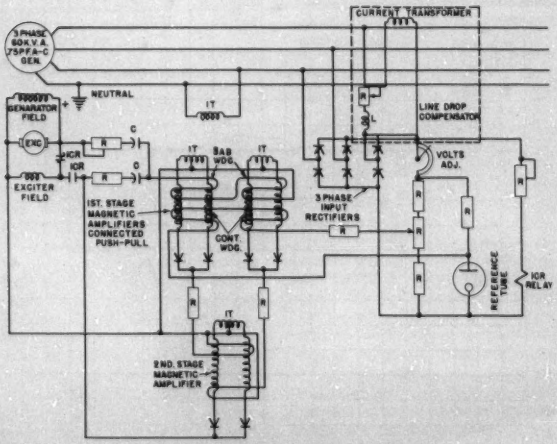


Fig. 7. Circuits of alternating voltage regulator used with 60-kva a-c generator

proportional to the synchronous reactance drop of the generator in magnitude and angle and the field current component I_s is proportional to generator output voltage; therefore, theoretically, the correct value of excitation required to maintain constant generator voltage will be supplied to the generator field.

A voltage regulator, however, is required on a system of this type to maintain the high degree of accuracy required. The system voltage is controlled by varying the current to the control winding of the saturable current transformer which effectively changes the value of X_m, which is in parallel with the generator field resistance R. By varying this impedance X_m, more or less current is shunted to or away from the field of the generator, so as to vary its terminal voltage. A voltage regulator of the type used on the previously described *NC-5* a-c equipment is employed to control this static excitation system.

Since this system is completely self-excited from the terminals of the a-c generator, it is usually necessary to "flash" the field of the a-c generator from the battery on the ground power unit, so as to initially energize the system. Once energized, the system will produce short-circuit current for all types of short circuits. This static excitation system produces rapid field forcing when starting induction motors, due to the generator load current feedback in the system. This system maintains voltage regulation of $\pm 1\%$, and the voltage recovers to its final value in approximately 0.1 second after application or removal of rated load.

A-C System Comparisons

The three a-c systems discussed included:

1. A semi-industrial type of a-c generator with an amplidyne exciter and a static magnetic-amplifier voltage regulator.
2. An a-c aircraft type of generator with a conventional d-c shunt exciter and a static magnetic-amplifier type of voltage regulator.
3. A semi-industrial type of a-c generator with a static exciter and a static voltage regulator.

If all of the a-c generators for these three systems had the same electrical constants, the over-all system electrical performance could be made comparable for all three systems with a somewhat faster response time being obtained from the static excitation system.

The semi-industrial type of system with the amplidyne exciter has the advantage over the aircraft-type a-c generator with the conventional d-c shunt exciter of longer life; however, it is the larger and heavier system. The economics and reliability of these two systems are comparable.

A static excitation system can be applied to a semi-industrial or to an aircraft type of generator. It presents the advantages of somewhat faster response and less maintenance because of no commutators. The static excitation system can be mounted in a location external to the generator which results in a shorter length for the rotating equipment and fewer mechanical problems. At present, the static excitation system usually costs somewhat more than the other two systems.

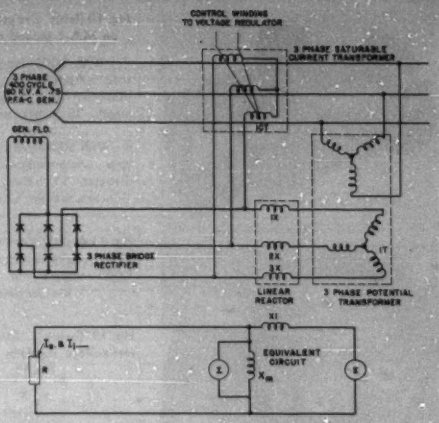


Fig. 9. Circuits of static excitation system used with 60-kva 400-cycle a-c generator

D-C Electric Systems

D-C Load Characteristics

In general, there are two separate classes of d-c loads on present military and commercial aircraft; these two classes are servicing loads and engine-starting loads.

D-c servicing loads include d-c motors, inverters, de-icing, lighting, and electronic equipment.

The military aircraft with d-c systems can be of two types, either the common-bus type, where the starters are connected to the main aircraft bus during starting, or the split-bus type where the starters are separated from the main bus during starting. The split-bus system of starting was set up so that the current to the starters could be controlled while the voltage on the main servicing bus could at the same time be maintained at a constant value.

It has been determined by considerable testing and maintenance experience that by properly regulating the current and voltage within close limits during the starting of the jet engines, minimum starting times may be obtained, and maintenance and failures of the aircraft starters and starter generators are minimized. The Air Force uses a system of applying 28 volts or under to the starter circuit initially and then varying the voltage to control the current after the starting cycle has been initiated. This system is known as a "hard" start. The Navy system applies approximately 6 volts to the starter motor initially in order to limit initial inrush current and torque on the starter. As the starting motor accelerates, the current is regulated and the applied voltage is increased to the regulated value. This system is known as a "soft" start.

General Performance Requirements For D-C Systems

The performance requirements for voltage regulation and recovery time on d-c ground power systems have changed very little over the past several years. Approximately $\pm 2\%$ voltage regulation and a 0.2-second recovery time have been adequate for most d-c applications. This voltage regulation and recovery time have been required over wide changes in temperature for the same reasons as mentioned previously in the section "General Performance Requirements For A-C Systems." D-c systems are often required to operate in parallel and to meet the same regulation and recovery requirements as are applicable for individual generator operation. When the

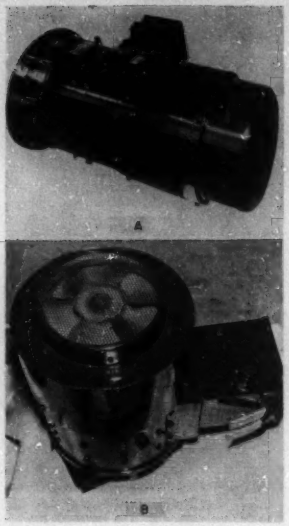


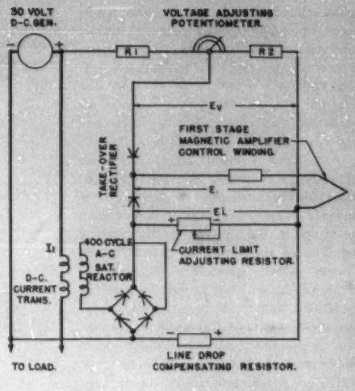
Fig. 10 (left). D-c generator used on MA-1 ground power unit

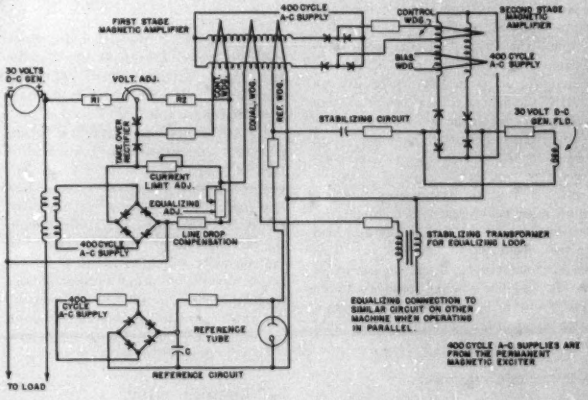
A—With 1,000-ampere main, 30 volts. Approximate dimensions: diameter, 15 inches; over-all length, 32 inches; weight, 660 pounds

B—With 500-ampere auxiliary, 30 volts. Approximate dimensions: diameter, 11 inches; length, 18 inches; weight, 235 pounds

Fig. 11 (right). Signal circuits of direct voltage regulator used with 1,000-ampere and 500-ampere d-c generator

Fig. 12 (below). Circuits of direct voltage regulator used as in Fig. 11





d-c ground power systems are required to start jet engines, the d-c generator output current is required to remain between 1,000 and 1,100 amperes during engine acceleration. Cable drop compensation is usually required on d-c systems so as to maintain the voltage regulation requirement at the airplane while large currents are flowing in the 30-foot interconnecting cable.

D-C System for MA-1 Ground Power Unit

The Air Force type-*MA-1* ground power unit employs two d-c generators for supplying d-c power. The main d-c generator is rated 30 volts direct current 1,000 amperes continuous, 1,500 amperes intermittent; and an auxiliary d-c generator is rated 30 volts direct current, 500 amperes continuous, 750 amperes intermittent. When operating in parallel, they have a combined rating of 1,500 amperes continuous, 2,250 amperes intermittent, at 30 volts direct current.

Static-magnetic regulators are employed to control these d-c generators. The use of static-magnetic amplifier components requiring an a-c supply is made possible on this system by employing a 3-phase 400-cycle permanent-magnet type of exciter, directly connected to the main d-c generator. This exciter has no brushes or commutator, so it is a very reliable device requiring a minimum of maintenance. A fixed voltage source of control power for switching relays and contactors is always available from this exciter even when the output voltage of the d-c generators varies during current regulation. The output of the ground power a-c generator could be used as the magnetic-amplifier supply for the d-c generator regulators; however, this makes the d-c system dependent on the a-c system.

Fig. 10(A) and (B) shows the two d-c generators with the permanent magnet exciter attached to the main generator. Contactor switching arrangement provides for either parallel operation of the two generators on the servicing bus, or for separate operation of the two generators with the main generator connected to the starting bus and the auxiliary generator connected to the servicing bus.

Fig. 11 is a simplified schematic diagram of the signal input circuit used in each voltage regulator. Voltage is sensed by a voltage divider *R1*, *R2*, and a voltage adjusting potentiometer connected across the terminals of the d-c generator. Current is sensed by use of a d-c transformer connected in series with the d-c generator output. This d-c transformer is essentially a saturable reactor with a one-turn d-c primary. For such a saturable reactor, the ampere turns in the secondary will be approximately equal to the current flowing in the one-turn primary. For any condition of current flowing in the primary, alternating current flowing in the secondary builds up to a point where one core unsaturates on each successive

half-cycle. Thus, the d-c transformer has rectified a-c output current proportional to the d-c output of the d-c generator. This is used to provide a signal for current regulation, for equalizing, and for line drop compensation. The voltage E which is the actual voltage supplied to the first-stage amplifier control winding is equal to a fraction of the d-c generator terminal voltage less a voltage drop across the line drop compensating resistor. This enables the voltage regulator to control a voltage equivalent to the voltage at the end of the cable. It thus regulates the generator terminal voltage in such a manner as to cause the voltage at the aircraft end of the output cable to remain constant.

The voltage E_1 is proportional to the current output of the generator and is measured across the current limit adjusting resistor. If the voltage E_1 is greater than the voltage E_2, the takeover rectifier enables the first-stage magnetic-amplifier control winding to regulate E_1 instead of E_2. The combined action of these sensing circuits is to regulate the generator voltage E_2 plus the cable drop, until it is overridden by the voltage signal proportional to the current E_1, and then to regulate the current output of the generator. The larger signal, representing either current or voltage, will be the regulated quantity.

Fig. 12 is a simplified schematic diagram of the regulating circuit for one d-c generator. It includes the signal input circuits already described, and a 2-stage magnetic amplifier to amplify the error signal and control the field voltage of the d-c generator to hold the regulated quantities constant. It also includes a reference circuit and an equalizing circuit for ensuring proper operation of the two generators in parallel when both are operating on the servicing bus. A constant reference voltage is obtained from the cold-cathode gas-filled reference tube. A current flows in the reference winding of the first-stage magnetic amplifier which is proportional to the constant reference voltage. A current from the signal input circuit flows in the control winding. The signal current flowing in the control winding of the first-stage magnetic amplifier opposes the reference current in the reference winding of the magnetic amplifier. If the ampere-turns due to the signal current become larger than the ampere turns due to the reference current, the first-stage magnetic amplifier's output voltage increases. This decreases the second-stage magnetic amplifier's output voltage, reducing the field current to the generator field and bringing

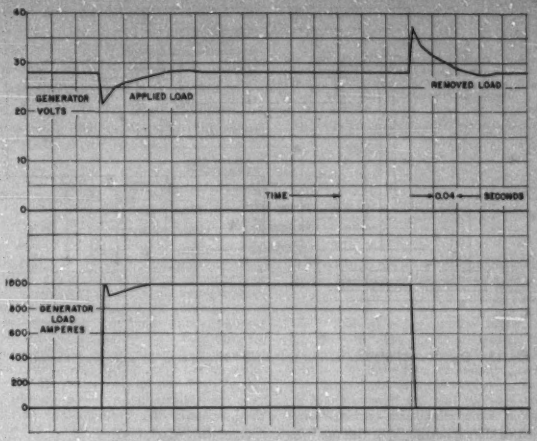


Fig. 13. Recording of d-c generator voltage and current upon application and removal of rated load

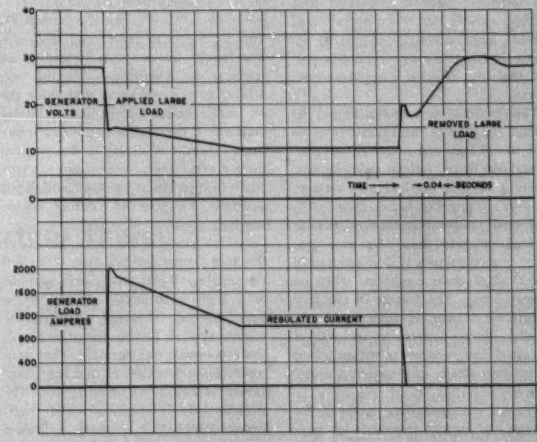


Fig. 14. Recording of d-c generator voltage and current upon application and removal of large load which causes current regulating circuits to reduce voltage

the generator voltage or current back to the regulated value.

An equalizing loop is provided through an equalizing winding. This loop measures the voltage on the equalizing resistor and compares it with a voltage on a similar equalizing resistor on the other regulator. If the current output from the generator shown in Fig. 12 is a larger proportion of its rating than that for the paralleled generator, its equalizing signal is such as to cause an equalizing current flow which decreases the output of the generator shown in Fig. 12 and increases the output of the generator with which it is operating in parallel.

It is necessary to provide two stabilizing circuits in this equipment; one stabilizing circuit degeneratively feeds back a voltage proportional to the rate of change of generator field voltage to the reference winding of the first-stage ampli-

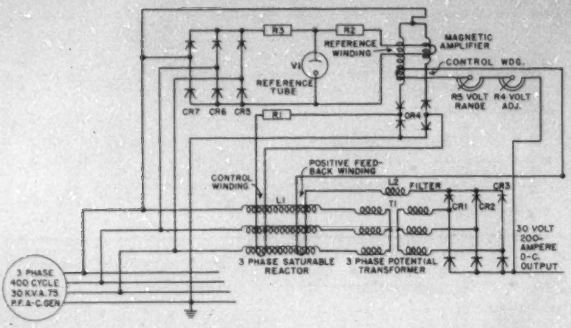


Fig. 15. Circuits of transformer rectifier unit rated 200 amperes, 30 volts direct current

fier. A second stabilizing circuit is provided in the equalizing loop through the action of a stabilizing transformer connected between the two generator fields when in parallel. This rate of change of the difference between generator field voltages ensures that when operating in parallel, large oscillations of current between the two generators will not persist, and that load division is rapid and accurate.

This electric system must maintain the voltage supplied to the aircraft from no load to full load and under all environmental conditions, without readjustment, within $\pm 2\frac{1}{2}\%$ of nominal rated voltage; however, it provides better than $\pm 1\%$ over small temperature changes between no load and rated load. Current regulation of minus 0 plus 10% is maintained under all of these conditions when on current limit. The voltage of the generators, when operating either singly or in parallel on application or removal of full rated load, recovers to within 5% of the ultimate steady-state value within 0.2 second.

Fig. 13 is a recording made of the voltage and current transients for each generator when applying and removing rated load.

Fig. 14 is a recording showing the application of a large load which causes the regulator to control current and reduce voltage.

In addition to the static magnetic voltage and current regulators, the d-c control equipment for the *MA-1* vehicle employs the following protective devices: positive polarity protection, reverse current protection for each generator, differential voltage protection, overvoltage and undervoltage protections, and overload protection for each generator.

Transformer Rectifier for NC-5 Ground Power Unit

For reasons of reliability, reduced maintenance, and flexibility of installation, it may be desirable to use transformer rectifier units in place of the d-c generators. The *NC-5* ground power unit utilizes a transformer rectifier unit instead of an auxiliary d-c generator. This unit also utilizes an a-c generator, and since the a-c loads required during engine starting are small, it is practical to utilize a transformer rectifier obtaining its supply from the a-c generator and supplying its output to the servicing bus, in place of the auxiliary d-c generator during engine starting.

Fig. 15 shows a schematic sketch of this transformer rectifier unit. The full a-c generator 3-phase voltage is supplied to a 3-phase saturable reactor whose impedance can be varied to control the voltage applied to the primary of a 3-phase potential transformer. The saturable reactor is called *L1* and the transformer is called *T1* in Fig. 15. The secondary of this 3-phase potential transformer supplies a 3-phase full-wave bridge-type rectifier *CR1*, *CR2*, and *CR3*, whose output is filtered by means of reactor *L2* before being applied to the cable going to the airplane. The rectifier output current also passes through a positive feedback winding on the saturable reactor *L1*, to lessen the power required from the voltage regulator to control the saturable reactor and also to give the system faster over-all response.

The voltage regulator is composed of a single-stage magnetic amplifier, with one control winding having a current proportional to the transformer-rectifier output voltage, and another winding having a constant current supplied by the voltage reference. The difference between these two currents is effective in controlling the magnetic-amplifier output which in turn controls the saturation in the 3-phase saturable reactor *L1*. The current in the control winding of saturable reactor *L1* controls the impendance of the saturable reactor, which in turn controls the voltage supplied to the primary of transformer *T1* in order to regulate the output voltage of the transformer-rectifier unit. A fan provides air to cool the output selenium rectifier in order to reduce the over-all size and weight.

This transformer rectifier unit is rated 200 amperes continuous, and 300 amperes for five minutes at 30 volts direct current. The output voltage is regulated within ± 0.75 volt between no load to full load and during changes of environmental conditions. The voltage recovers to its final value in less than 0.2 second following changes between no load and rated load.

D-C System Comparisons

The transformer-rectifier unit as a source of direct current offers the advantage over the rotating d-c generator of requiring less maintenance. The transformer-rectifier unit involves fewer mechanical problems, provided a source of alternating current is already available. If the source of a-c power is 400 cycles, the performance of the two systems can be comparable. The efficiency of the transformer-rectifier unit is usually less than that of the rotating d-c machine (especially if selenium rectifiers are used) and it is difficult to obtain low-ripple voltage at light load in the transformer-rectifier unit without sacrificing efficiency, size, and weight.

Conclusions

Power Requirements

The recent trend in ground power is toward more and larger a-c equipments since most modern airplanes employ a-c systems. The trend is also toward less d-c ground power equipment because few modern airplanes employ d-c systems and because engine starts are being performed by compressed air and combustion instead of d-c motors.

Performance Requirements

The performance requirements are becoming progressively more rigid because the ground power is being used as a reference for calibrating complex electronic equipment.

PHYSICAL REQUIREMENTS

The recent trend is for equipment which offers greater reliability, longer life, and which requires less maintenance consistent with small size and light weight. Excessive temperature rises in the rotating equipment are being avoided. The aircraft type of control devices such as relays and contactors are usually employed. The static magnetic-amplifier type of regulating equipment is fast replacing the electronic and electromechanical types. It is expected that static-type systems will continue to replace d-c generators and d-c exciters, thus lessening maintenance problems.

References

1. THE DEVELOPMENT OF A STATIC VOLTAGE REGULATOR FOR AIRCRAFT A-C GENERATORS, H. H. Britten, D. L. Piatte. *AIEE Transactions*, vol. 74, pt. II, 1955 (Jan. 1956 section), pp. 438–43.

2. STATIC MAGNETIC EXCITER FOR SYNCHRONOUS ALTERNATORS, H. F. Storm. *Ibid.*, vol. 70, pt. I, 1951, pp. 1014–17.

Discussion

Erle P. Barlow (The Glenn L. Martin Company, Baltimore, Md.): The authors' reasons appear valid for the use of the ground power supply indicated. Engineers who have experienced equipment malfunction in aircraft, only to discover that the equipment functions nicely when supplied with precise sine wave power of optimum characteristics, are naturally concerned with the possibility that the existence of ground power supplies, having characteristics so favorable to equipment operation, will mislead the equipment manufacturers into an expectation of similarly desirable characteristics in the aircraft power system. In short, air-borne electronic equipment should be designed to function satisfactorily with realistic air vehicle power systems.

The electronics equipment manufacturer is, of course, at liberty to use any quality of power he wishes during the development of his equipment; but before this equipment can be considered properly qualified for production or service acceptance, it should be demonstrated as meeting the most unfavorable characteristics permitted by the applicable power system specification; e.g., for military application, specification *MIL-E-7894*. Assuming that this equipment capability has been established and has been maintained in production by suitable quality control, the authors' point of view that subsequent ground calibration of the equipment can be more quickly and accurately done on a precise power supply is undoubtedly valid. The final measure of the equipment's effectiveness, however, will remain its ability to perform its function satisfactorily on air-borne power systems.

H. S. Sechrist: The authors concur with Mr. Barlow. The nature of complex electronic equipment evidently makes parts of it sensitive to wave form, modulation, recovery time, etc., of the a-c supply voltage and, as Mr. Barlow points out, the only absolutely positive test is for this electronic equipment to meet its specified performance with the plane's a-c supply. It is not presently considered economically feasible to have a separate ground power unit for each airplane consisting of the same electrical generating and regulating equipment as that of the particular plane; therefore, the best compromise appears to be the use of precise ground power.

Design of a Contactor Servo Using Describing Function Theory

STANLEY K. CHAO
ASSOCIATE MEMBER AIEE

Synopsis: The theory of describing function method used in analyzing and synthesizing nonlinear control systems has been supported so far largely by practical systems simulated in analogue computers. This paper presents a practical contactor servo system which is designed on the basis of this theory. The experimental result checks with theoretical prediction within engineering accuracy. The practical validity of this theory is thus further verified.

CONTACTOR servomechanisms have been used and studied for a number of years. They are widely used because the contactor contained in the system provides large gain which might otherwise require elaborate amplification devices. A contactor servomechanism can be designed with very few components and which is probably free of troublesome electronics. Theoretical study of contactor servo systems has not been fully developed in pace with practical uses mainly because of the complex mathematics involved in the analysis of discontinuous system. Goldfarb[1] in 1947 and Kochenburger[2] in 1949 independently proposed the describing function method (also known as frequency response method) for analyzing and synthesizing nonlinear systems. In the same direction, Johnson[3] gave a rigorous derivation showing the validity of applying the describing function method to nonlinear servomechanisms and the inaccuracies that can be expected in employing this method. Describing functions for hysteresis, dead space, negative-deficiency, and saturation have also been derived by Greif.[4]

Applicability of this method to practical design problems has been demonstrated so far largely through the help of computers where a "designed" servo system is simulated and its performance computed. In applying the theory to design problems, engineers are usually faced with practical situations not quite so idealized as those in a computer. Then just how practical is the theory? How large a discrepancy is to be expected between their theoretical and practical results? And what are the best design procedures to be followed? This

Paper 56-456, recommended by the AIEE Feedback Control Systems Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE-ASME Nonlinear Controls Conference, Princeton, N. J., March 26–28, 1956. Manuscript submitted October 11, 1955; made available for printing February 24, 1956.

STANLEY K. CHAO is with the Atomic Instrument Company, Cambridge, Mass.

The material presented in this paper is based upon the results of a research project undertaken during 1953–1954 in the Feedback Control Laboratory of the Massachusetts Institute of Technology, Cambridge, Mass. The author wishes to acknowledge with sincere gratitude the many helpful suggestions offered by Prof. Paul E. Smith, Jr., in his capacity as project supervisor.

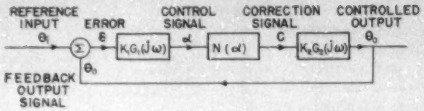


Fig 1. (above). Typical form of contactor servomechanism

Fig. 3 (right). Plot of $-\frac{1}{N(\alpha)}$ for contactor

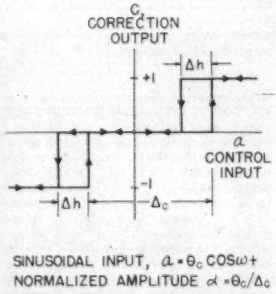
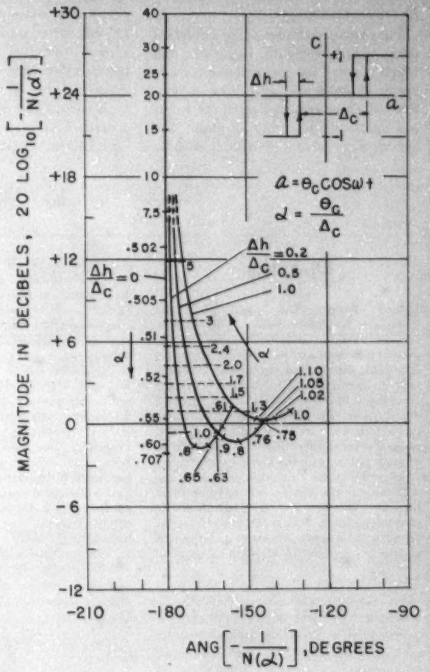


Fig. 2. Contactor with hysteresis and dead zone, dimensionless representation



paper presents a contactor servo whose design is based on the theory of describing function method. The system performance checks fairly well with the predicted result. Thus, the practical applicability of the method is established. Design techniques based on this method are clearly demonstrated for the simple contactor servo system considered.

Normally, when a contactor servo is designed, the describing function theory would be used to predict the system behavior. And in synthesizing the system, it is used to show how sustained oscillation could be eliminated or reduced. But here, in order to confirm the theory's prediction of sustained oscillations and to verify the predicted amplitude and frequency of such oscillations, an oscillatory servo was designed and its behavior investigated. This paper is a report of the result of this investigation.

Describing Function Theory for Contactor Servo

To facilitate concrete and complete presentation, the describing function theory is first reviewed.

Usually a servo containing a single contactor as the only nonlinear element may be represented by a block diagram, as shown in Fig. 1. The symbols used are:

K_1 = static gain of linear compensating network
$G_1(j\omega)$ = transfer function of linear compensating network
$N(\alpha)$ = describing function of contactor, frequency independent
α = normalized amplitude of input signal to contactor
K_2 = static gain of output member
$G_2(j\omega)$ = transfer function of output member

The theory of describing function is based on two assumptions: first, the control signal α, impressed on the contactor, is sinusoidal. Second, the resultant response of the contactor is periodic and can be satisfactorily represented by its fundamental Fourier component in so far as the over-all system performance is concerned. Thus the describing function of a contactor is defined as the amplitude and phase angle of the fundamental component of the response with respect to the assumed impressed sinusoid. To study a system's behavior, describing function for the contactor may be combined with, and treated the same way as, the transfer function of the linear portion of the system. In this way, sinusoidal analysis may be applied.[2]

Following this deduction, the over-all transfer characteristics of the system in Fig. 1 can be written as

$$\frac{\theta_o}{\theta_i} = \frac{K_1G_1(j\omega)N(\alpha)K_2G_2(j\omega)}{1+K_1G_1(j\omega)N(\alpha)K_2G_2(j\omega)} \quad (1)$$

If $G(j\omega)$ is defined as the linear portion of the composite transfer function thus

$$G(j\omega) = K_1K_2G_1(j\omega)G_2(j\omega) \quad (2)$$

and equation 1 becomes

$$\frac{\theta_o}{\theta_i} = \frac{G(j\omega)N(\alpha)}{1+G(j\omega)N(\alpha)} \quad (3)$$

The system will become unstable and oscillation may be self-sustained when

$$1+G(j\omega)N(\alpha)=0 \quad (4)$$

or when

$$G(j\omega) = -\frac{1}{N(\alpha)} \quad (5)$$

Investigation of equation 5 can be best performed by a graphical method. If the curve representing $G(j\omega)$ intersects with the curve representing $-1/N(\alpha)$, equation 5 is satisfied and the system will oscillate. The amplitude α and the fre-

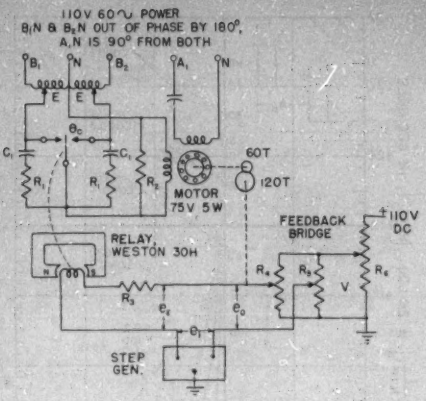


Fig. 4 (left). Schematic diagram: single contactor servo

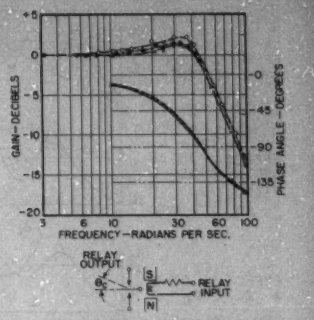


Fig. 6. Normalized relay circuit frequency response, $G_{RC}(j\omega)$. Voltage on contact for curves as follows:

X—25 volts
●—52 volts
○—75 volts

quency ω of the oscillation are determined by the intersecting point. To facilitate rapid synthesis, $G(j\omega)$ is presented on gain-phase plane. The inversed describing function $-1/N(a)$ should also be plotted on the gain-phase plane so that it can be superposed on the curve for $G(j\omega)$.

A contactor with hysteresis and dead zone has output-to-input relationship, as shown in Fig. 2. The describing function for this contactor has been derived and plotted on a dimensionless linear plane by Kochenburger.[2] His result is replotted on gain-phase plane in Fig. 3.

Single Contactor Servo System

It is desired that a servo system containing one contactor as the nonlinear element be designed to demonstrate the practicability of describing function theory. For simplicity and ruggedness, it should consist of the least number of components and be free of vacuum tubes. The components used should be of common commercial items; thus specially designed parts will be avoided. Coulomb friction, contact arcing, and other nonlinearities usually inherent in a contactor servo are neglected in the describing function just stated. In a practical design, they must be reduced to the negligible amount or else the describing function must be modified to include them.

Circuit

Fig. 4 shows the single contactor servo system. The contactor selected is a moving coil type of polarized sensitive relay. Its contact-spacings at both sides can be adjusted at will to give the desired value of dead zone. The relay is used not only as the nonlinear contactor but also as an error-sensing element. Its coil is energized by an error voltage which is the difference between the step input from a signal generator and the output from the output member.

When relay contacts carrying current are opened and closed, there exist the phenomena of contact arcing and contact bouncing.[5] Both are severely destructive to the relay contacts when the load is direct current. In the case of a-c load, the danger is greatly lessened because the arc will be automatically extinguished when the sinusoidal current passes zero point. Nevertheless, when the contacts are used in a servo system, having sustained oscillation such as the case here, arcing, though short in its duration, is frequent enough to be excessive. Contact protecting network is thus used. It consists of one resistance-capacitance network on each pair of contacts, and R_2. The latter is connected across the control field of the two-phase motor to absorb part of the inductive energy when contact is opened and to help reduce the size of capacitor C_1.

The selection of a two-phase motor as the output motor is arbitrary and it could as well be a series or a shunt motor. Gear train is reduced to two gears in order to minimize the coulomb friction and backlash.

A "feedback bridge" is used to convert the output motor shaft motion into electric signal so that it can be compared with the input. Potentiometer R_4 is a high-power wire-wound type and its armature spring force is reduced to mini-

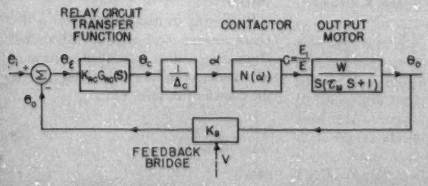


Fig. 5. Block diagram of servo system in Fig. 4

Table I. Motor* Runaway Velocity and Time Constant

Reference Field Voltage, Volts	Control Field Voltage, Volts	Runaway Velocity,† W, Radians per Second	Time Constant, τM, Seconds
110	25	57.0	0.0662
110	52	82.0	0.0425
110	75	88.5	0.0377
75	25	65.0	0.0638
75	52	84.0	0.0571
75	75	88.0	0.0332

* Two-phase motor, 75 volts 5 watts, Delhi-*FPE*.
† Including gear ratio, see equation 13.

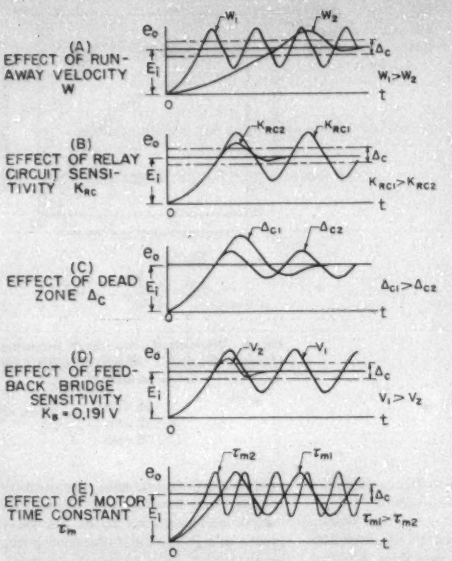


Fig. 7 (above). Effect of parameter magnitude on system response

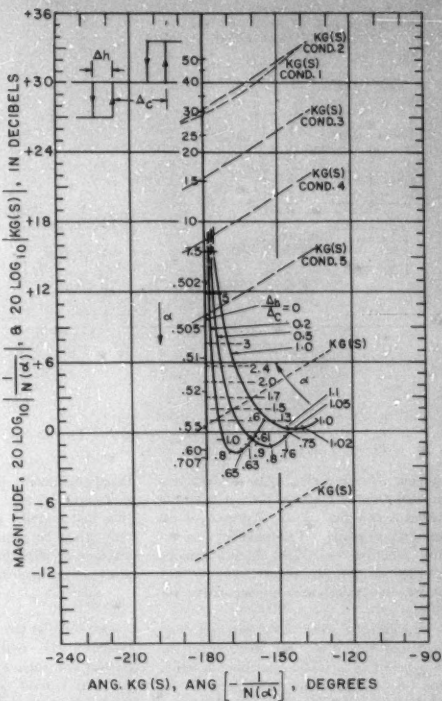


Fig. 8 (right). Gain-phase curves of servo, linear and nonlinear portions

mize friction. Adjustment of potentiometer R_3 shifts the center position of R_1, about which its arm will oscillate. Potentiometer R_4 controls the bridge sensitivity and thus the feedback gain of the system.

Description of Parameters

In Fig. 4 the error voltage e_ϵ will appear across the relay coil including R_2 whenever output e_0 is different from input e_i. The current resulting from this error voltage will produce a magnetomotive force causing the center arm of the contactor to move correspondingly both in polarity and in magnitude. The angular movement of this arm θ_C is therefore proportional to the magnitude and sign of the error voltage. This relationship can be represented by

$$\frac{\theta_C}{e_\epsilon} = K_{RC} G_{RC}(s) \tag{6}$$

where

Table II. Various Stability Conditions

Condition	V_C, Volts	V, Volts	W, Radians per Second	τ_m, Seconds	ρ	ω_n Radians Per Second	K, Decibels
1	52	110	82.0	0.043	0.42	42.6	57.6
2	75	110	88.5	0.038	0.37	42.6	58.3
3	75	55	88.5	0.038	0.37	42.6	52.2
4	75	30	88.5	0.038	0.37	42.6	46.9
5	75	15	88.5	0.038	0.37	42.6	40.5

V_C = Motor control field voltage (reference field at 110 volts for all cases)
V = bridge voltage
W = runaway velocity
τ_m = motor time constant
ρ = relay circuit damping ratio
ω_n = relay circuit resonant frequency
K = loop gain

K_{RC} = relay circuit static gain, radians per volt
$G_{RC}(s)$ = relay circuit transfer function, nondimensional

If the nonlinear contactor itself is tentatively represented by $K_N G_N$, it follows that

$$\frac{E_1}{\theta_C} = K_N G_N \tag{7}$$

where

E_1 = fundamental component of the contactor output, volts
θ_C = contactor arm movement, directly related to relay coil current, radians

and both K_N and G_N are frequency invariant.

Neglecting its coulomb friction, a two-phase servomotor, having its control field alternating voltage controlled by forward-off-reverse operation, is represented by the transfer function

$$\frac{\theta_m(s)}{E_1} = \frac{K_m}{s(\tau_m s + 1)} \tag{8}$$

where

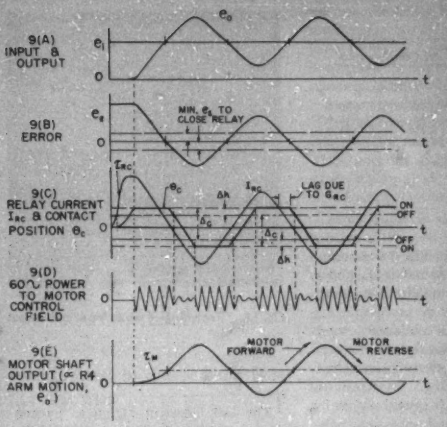


Fig. 9 (above). Temporal analysis of servo system in Fig. 4

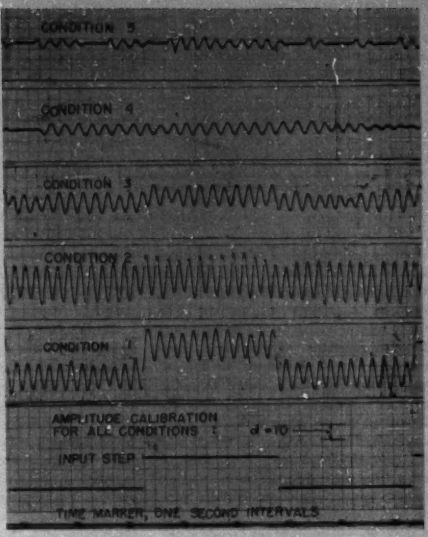


Fig. 10 (right). Recordings of servo output

θ_m = motor output, radians
E_1 = input to motor control field
= fundamental component of the envelope of contactor output in terms of voltage, volts
K_m = motor constant, radians per second per volt
= $(\delta n/\delta I_m)(\delta I_m/\delta E_1)$ (see pages 131-33, reference 6)
n = motor speed, radians per second
I_m = control field current, amperes
τ_m = motor time constant, seconds

When gear ratio is defined as the reduction ratio between the motor shaft position (θ_m) and the output shaft position (θ_0), then

$$\frac{\theta_0}{\theta_m} = \frac{1}{g} \tag{9}$$

The open-loop transfer function may now be formulated by combining equations 6 through 9 which obtains the following equation

$$\frac{\theta_0}{e_e} = K_{RC}G_{RC}(s) \times K_N G_N \times \frac{K_m}{s(\tau_m s+1)} \times \frac{1}{g} \tag{10}$$

To make the function dimensionless, rewrite the product of operators constituting θ_0/e_e as follows

$$\frac{\theta_0}{e_e} = K_{RC}G_{RC}(s) \times \frac{1}{\Delta_C} \times (\Delta_C K_N G_N)\frac{1}{E} \times \frac{\frac{EK_m}{g}}{s(\tau_m s+1)} \tag{11}$$

where

Δ_C = contactor dead zone, radians
E = line voltage to control field of motor, volts

The third factor is in the form of the dimensionless describing function as presented in Fig. 3, since, from equation 7 is obtained the following

$$(\Delta_C K_N G_N)\frac{1}{\beta} = \frac{\frac{E_1}{E}}{\frac{\theta_C}{\Delta_C}} = N(\alpha) \tag{12}$$

where

$\alpha = \frac{\theta_C}{\Delta_C}$ = nondimensional input to contactor

The frequency invariant portion of the last factor in equation 11 is defined as the runaway velocity

$$W = \frac{EK_m}{g} \text{ radians per second} \tag{13}$$

Runaway velocity is directly proportional to the system speed of response. It is desirable, within stability limitations, to have as large a runaway velocity as possible so that the error following some disturbance will be reduced to a tolerable level at the shortest possible time. Using these representations, block diagram for the system in Fig. 4 has been drawn in Fig. 5.

Determination of Parameter Values

Expressions and values for the symbols in the block diagram will now be sought. To get closer representation of the components used, whenever possible, expressions or values are determined through actual test rather than through the mathematical expression obtained from the manufacturer's published data. The relay circuit static gain K_{RC} was measured by first setting the contact openings between the center arm and the contacts of the relay at a given value and then applying an adjustable direct voltage until the arm touched the contact; see Fig. 4. Since the relay arm moves on a circular arc, its separation from the contacts is expressed in radians. This static gain was found to be

$$K_{RC} = 0.011 \text{ radian per volt} \tag{14}$$

In this measurement, the contactor was made to carry the operating voltage of the motor field. This step was taken as a precaution to account for the possible effect of the electrostatic attraction under operating conditions. However, no such effect was detected.

The magnitude of the transfer function of the relay circuit was measured in a similar manner. The standard frequency response-measuring technique is to measure the response for a given input. But here it was found convenient to set the

Table III. Summary of Results

Condition	Gain, Decibels	Theoretical Result f	Theoretical Result α	Experimental Result f	Experimental Result α	Per Cent Deviation f	Per Cent Deviation α
1	37.6	4.45	28.6	4.75	20.4	−6.75	+28.1
2	58.3	4.54	30.0	4.85	25.4	−6.85	+15.0
3	52.2	4.47	15.5	4.30	13.8	+3.80	+11.0
4	46.9	4.47	8.4	4.16	7.9	+6.95	+ 5.3
5	40.5	4.44	4.0	*	*	*	*

* = no sustained oscillation observed for this condition.
$f = \omega/2\pi$ = oscillation frequency in cycles per second.
α = nondimensional amplitude of oscillation.

relay circuit output (the contact opening) at the specified value and then to measure the input required at various frequencies to produce this output. The input was provided by a low-frequency function generator. The result of these measurements is presented in Fig. 6, examination of which shows that the magnitude frequency response of the relay circuit varies slightly with the voltage carried by the contacts at the region near the resonant frequency. Regardless of the voltage applied to the contacts, relay circuit transfer function exhibits zero decibel magnitude at low frequencies. A slope of −40 decibels per decade is observed in all cases for higher frequency range. At resonant frequency their magnitudes vary from +1 to +2 decibel. Therefore, the transfer functions resemble the quadratic form

$$K_{RC}G_{RC}(s) = \frac{K_{RC}}{\left(\frac{s}{\omega_n}\right)^2 + 2\rho\frac{s}{\omega_n} + 1} \qquad (15)$$

where ω_n is the resonant frequency and ρ the damping ratio. The variations of ρ and ω_n for various contact voltages is caused by electrostatic attraction and contact arcing. When a higher voltage is applied across the contacts, the attracting force between the contacts is increased. As the contacts are approaching each other, this force will accelerate and finally pull them closed. Thus for a given contact opening, the effective result of a higher voltage across the contacts is a decrease of input voltage to relay coil for the contacts to close. This is equivalent to an increase of contactor gain. Arcing between contacts has the same but less pronounced effect because of the effectiveness of arc suppressors used.

Since the theoretical behavior of the moving coil relay is that of a second-order system, equation 15 is complete in describing the behavior of the relay circuit. If equation 15 completely describes its behavior, the knowledge of the parameters ω_n, ρ, and K_{RC} specifies not only the amplitude but also the phase angle of the frequency response of the system. The phase angles shown in Fig. 6 were calculated from the corresponding ω_n and ρ, obtained from their amplitude measurements.

Dead zone Δ_C of the contactor depends primarily on the separation of contacts. It is also affected by electrostatic attraction and contact arcing. Hysteresis Δ_h in this sensitive relay is caused mainly by attraction, arcing, and vibration. Detailed discussion on the sources, the effects, and the control of both Δ_C and Δ_h is given in Appendix I. For an oscillatory system, it is necessary to have small Δ_C, large Δ_h and, consequently, a relatively large value of Δ_h/Δ_C. In the system designed, dead zone was set at

$$\Delta_C = 0.025 \text{ radian} \qquad (16)$$

or an equivalent of 345 microamperes in the relay coil. Hysteresis, being dependent on arcing, is very difficult to control. If arcing is unsuppressed, Δ_h, will vary widely for successive operations, depending on which part of the a-c power cycle is being interrupted at the contacts. This difficulty and the necessity of protecting the life of the contacts made it mandatory to reduce arcing to a practical minimum. The "residual" arcing plus the effect of the electrostatic force causes existence of Δ_h. It was measured to be

$$\Delta_h = 0.004 \text{ radian} \qquad (17)$$

and was 55 microamperes in terms of relay coil current.

From equations 16 and 17

$$\frac{\Delta_h}{\Delta_C} = 0.16 \qquad (18)$$

Runaway velocity of the output motor was measured and tabulated in Table I for different control field voltages. The reference field was overexcited at 110 volts. Motor time constant was measured with the use of a method described in Appendix II. Note that motor time constant also varies with the control voltage used. This is consistent with the speed torque characteristics of the motor, since $\tau_m = J(\delta n/\delta T)$, and the slope

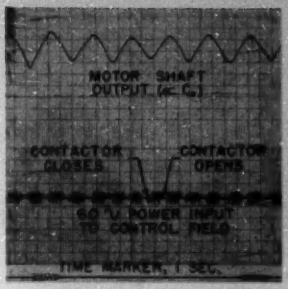


Fig. 11. Recording of motor input and output

$\delta n/\delta T$ varies slightly at different motor speed.

The last block in Fig. 5 to be considered is the feedback constant K_B. It is the sensitivity of the feedback bridge. When potentiometer R_4, Fig. 4, has a maximum angle of rotation of 300 degrees, the feedback constant is

$$K_B = -\frac{V}{2\pi\frac{300}{360}} = 0.191V, \text{ volts per radian} \qquad (19)$$

where V is the direct voltage applied to the bridge.

The over-all open-loop gain of Fig. 5 can now be written

$$\frac{e_0}{e_i}(s) = K_{RC}G_{RC}(s) \times \frac{1}{\Delta_C} \times N(\alpha) \times \frac{W}{s(\tau_m s + 1)} \times K_B \qquad (20)$$

The linear portion of equation 20 can be combined as

$$KG(s) = K_{RC}\frac{1}{\Delta_C}WK_B\frac{G_{RC}(s)}{s(\tau_m s + 1)} = K\frac{G_{RC}(s)}{s(\tau_m s + 1)} \qquad (21)$$

The loop gain in equation 21 can be rewritten as

$$K = \frac{1}{\Delta_C} \times W \times K_{RC}K_B \qquad (22)$$

The first factor $1/\Delta_C$ is the gain in the contactor itself. The choice of its value predetermines the size of the error voltage necessary for oscillatory conditions. The second factor, the runaway velocity W, is the gain between the contactor and the output motor. Its value determines not only the speed of response, but also the magnitude of the torque produced by the motor in overcoming coulomb friction. The last factor is the gain preceding the

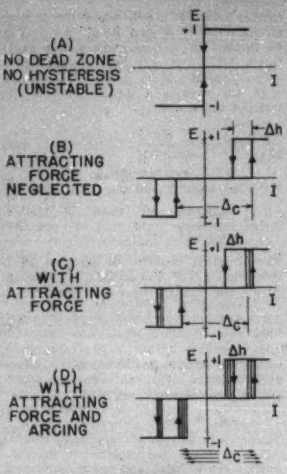


Fig. 12. Effect of contact-voltage on hysteresis and dead zone

contactor. This gain must be large enough to cause the relay contacts to close. To summarize the effect of each parameter in equation 21, Fig. 7 gives sketches of the system response to a step input for different magnitude of each parameter while holding all other parameters constant. For good stability, the runaway velocity should be small, but the speed of response is then reduced. Both K_{RC} and K_s should be small and Δ_C should be large for stable operation, but they all tend to increase the region of possible steady-state error. Again for stability considerations, the motor should have a large time constant, but it impairs the speed of response. Therefore these parameters should all be carefully considered and an optimum set of parameters chosen in meeting design specifications.

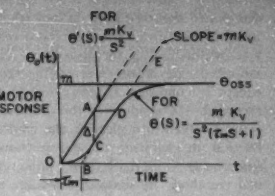
Graphical Determination of System Operation

For comparison, five stability conditions are assumed. They are to have different gain adjustments as well as different contact-voltage (voltage which is carried by the contacts and subsequently applied to the motor control field). Table II is a tabulation of the conditions assumed. Condition 2 is the same as condition 1 except the voltage carried by the contacts; the former has 52 volts whereas the latter has 75, which is the rated value of the motor control field. From condition 3 on, the servo loop gain is successively decreased through the adjustment of the bridge voltage V.

As previously described, the values of W, τ_m, ρ and ω_n are in one way or another related to the motor control field voltage. Thus, once the value of V_c is chosen in each condition, all other values can be immediately specified.

From Table II and Fig. 6, the linear portion of the open-loop transfer function set forth in equation 21 can be calculated and plotted on the gain-phase plane. In Fig. 8 the $KG(s)$ of all five conditions are superposed on the plot of $-1/N(\alpha)$. Intersections of the linear and the nonlinear functions give the amplitude and frequency of oscillation of the contactor servo.

Temporal behavior of the designed system is now analyzed. Fig. 9(A) shows the step input and the oscillatory output. Output wave form is very nearly sinusoidal because of the low-pass-filter action of the servomotor. Error signal is shown in Fig. 9(B). There is a minimum value of error signal required to close the relay contacts. This value is equal to $(\Delta_C+\Delta_h)/2$. When an error voltage appears at the relay circuit, the resultant current is shifted slightly in phase due to the time constant $\tau_{RC}=L_{RC}/R_{RC}$. This time constant has been neglected in previous study because R_{RC}, consisting of the relay coil resistance as well as R_s, is much larger than the relay coil inductance L_{RC}. As the relay circuit current changes, the contactor arm will move from its off-center position correspondingly. But due to G_{RC}, the arm motion θ_c would lag behind the actuating current I_{RC}. This relationship is shown in Fig. 9(C). Power is applied to the output motor control field as soon as the contact arm makes with either contact. The number of cycles of a-c power per contact closure is determined by the period of oscillation of the system. Equal time interval is assumed for both forward and reverse relay closures. Irregularities due to arcing and electrostatic attraction are to be expected. Transient due to a sudden interruption of a-c power is neglected in Fig. 9(D). Subject to its own time constant τ_m, the motor starts to rotate potentiometer R_4 when its control field is energized. This gives rise to the output voltage e_0. Note in Fig. 9(E) that the sinusoidal wave form describing the motor shaft rotation is not symmetrical with respect to its stationary position. The amount of displacement is adjustable through potentiometer R_5.



θoss = STEADY STATE RESPONSE
m = STEP SIZE
Kv = MOTOR VELOCITY CONST.
τm = MOTOR TIME CONSTANT

Fig. 13. Determination of motor time constant

Experimental Results

Both the input and the output of the servo system were recorded on a Sanborn oscillographic recorder, one on each channel. The frequency of oscillation was apparent since the paper speed was known. The amplitude of oscillation a was obtained by normalizing the recorded amplitude by the equivalent of the dead zone Δ_C. Recordings of various conditions are displayed in Fig. 10.

Condition 1 had a lower motor control field voltage of 52 volts, down 30% (per cent) from the rated value of 75 volts. The setting of this lower voltage was intended to show the effect of the magnitude of voltage being carried by the contacts. The magnitude of this voltage affects the contactor dead zone Δ_C, the hysteresis Δ_h, th relay circuit transfer function $G_{RC}(s)$, the motor time constant τ_m, and the motor runaway velocity W. These various effects are demonstrated in Figs. 6 and 12, and Tables I and II.

Conditions 2 through 5 had the rated motor control field voltage, but their bridge voltages were successively decreased. No sustained oscillation was observed when the bridge voltage was decreased to 15 volts in condition 5. While the motor fields were excited with the rated voltage, this gave a fair indication as to the duration of field excitation, Fig. 9(C) and (D), required to produce enough output power in overcoming the coulomb friction in the system. Coulomb friction had to be overcome in order to bring the contactor out of its dead zone at each time so as to sustain oscillation. Apparently coulomb friction was not overcome in condition 5. Oscillation, though occasionally started, did not sustain.

The theoretical and experimental results of the frequency and amplitude of oscillation for each condition are summarized in Table III. When oscillation

did exist, the deviation of the experimental result from the theoretical result was within ±7% for frequency. The amplitude deviation was within 15%, except in condition 1, where 28.1% was recorded. Considering the inevitable error involved in neglecting the coulomb friction, in idealizing the motor transfer function as $K_m/s(\tau_m s+1)$, in using graphical method in data recording and theoretical calculation, and in neglecting the harmonics in contactor describing function, the results are regarded as fairly good, and within engineering accuracy.

In Fig. 11, a slow-speed recording was taken to show the motor output as well as the 60-cycle power input to the motor control field. As expected, the duration in which the 60-cycle input signal is applied varies irregularly due to contact arcing during each relay closure.

Conclusions

From the investigation described, the following conclusions can be drawn. Describing function theory as derived by Kochenburger neglects coulomb friction effect which exists in all practical servo systems. It is therefore applicable only to contactor servo where coulomb friction is small compared with the motor torque. Discrepancy between this theory and practical system depends on the relative magnitude of coulomb friction with respect to the motor torque. The higher the friction, the larger the discrepancy. The discrepancy can be reduced by using larger output motor to belittle the influence of friction effect. For conditions where larger motor is not advantageous, it is suggested that the contactor describing function be revised to include the effect of coulomb friction.

The voltage carried by the contacts of the relay in a contactor servo is of vital importance to the performance of the system. It influences directly the value of contactor dead zone, contactor hysteresis, and contactor circuit transfer function; and, indirectly, if output motor is immediately following the contactor, the motor time constant, and the motor runaway velocity. This voltage should therefore be carefully considered in designing contactor servo.

Appendix I. Hysteresis and Dead Zone in Contactor Servomechanism

A relay will remain inactive until a sufficient current is applied to its coil. Thus dead zone is inherent in a relay. To insure positive contact closing, the relay pull-in current is necessarily greater than its drop-out current. This constitutes hysteresis. However, since the contacts are carrying current and usually have a voltage applied across them when open, both dead zone and hysteresis become complicated. Electrostatic attraction between the contacts when they are very close to each other will tend to close the contacts even though the current in the coil is insufficient to do so. While it depends on the magnitude of voltage, the average attracting force is greater for direct current than it is for alternating current. The effect of this force is a decrease in the amount of coil current needed to close the contact (thus a reduction in dead zone). For a continuously on-and-off relay operation, there is no reason to expect this force to produce exactly the same amount of reduction in dead zone for each operation.

When a relay is being de-energized, arcing due to the interruption of current carried in contacts contributes greatly to the effective hysteresis. The contacts may not be physically in contact, but the arc continues to carry load current. If an alternating current is carried, the magnitude of arcing depends upon which point of the sinusoidal wave form the current is being interrupted. If a direct current is carried, the amount of hysteresis contributed by arcing is also markedly irregular from one operation to the next.

In general, the presence of electrostatic attraction decreases dead zone and the existence of contact arcing increases hysteresis. Thus in the describing function, the factor Δ_h/Δ_C is increased. Examination of Fig. 3 shows that for a contactor servo experiencing sustained oscillation, an increase in Δ_h/Δ_C decreases the frequency and increases the amplitude of oscillation. In addition, the increase in Δ_h/Δ_C is not constant but variable in an unpredictable manner, as demonstrated in Fig. 12.

Therefore, in designing a contactor servo system where stability is a prime consideration: 1. the differential between relay pull-in and relay drop-out currents should be as small as possible, 2. the hysteresis due to contact arcing should be reduced to a minimum by using arc suppressing network, 3. the contactor voltage should be at minimum consistent with the power requirement at the output member, 4. the relay dead zone should be as large as practical and still within the limitation of specified tolerable static error which exists before the output member is caused to move so as to reduce the error, and 5. contact bounce and the influence of external mechanical vibration should be minimized by optimum relay design.

Appendix II. Graphical Method of Determining Motor Time Constant

The method to be described is used to determine experimentally the time constant in a 2-phase motor. This method can be used to determine the time constant of any system where the transfer function contains an integration and a first-order factor, $1/(TS+1)$, and where the system is subjected to a step input. Or it can be applied to systems with a first order factor $1/(TS+1)$ only but subjected to a velocity input.

In the case of a 2-phase motor, the transfer function is

$$\frac{\theta(s)}{E(s)}=G(s)=\frac{K_g}{s(\tau_m s+1)} \tag{23}$$

For a step input of m in magnitude, $E(s)=\frac{m}{s}$

$$\theta(s)=\frac{mK_g}{s^2(\tau_m s+1)} \tag{24}$$

The transient response of equation 24 is shown in Fig. 13. For an ideal motor where $\tau_m=0$, the response to a step input is

$$\theta'(s)=\frac{mK_g}{s^2} \tag{25}$$

which is a straight line in the $\theta_0(t)-t$ plane as shown in Fig. 13. The deviation between equations 24 and 25 is

$$\Delta(s)=\frac{mK_g}{s^2}-\frac{mK_g}{s^2(\tau_m s+1)}=\frac{mK_g\tau_m}{s(\tau_m s+1)} \tag{26}$$

since $\Delta(t)=\mathcal{L}^{-1}\Delta(s)$, and using final value theorem

$$\lim_{t\to\infty}\Delta(t)=\lim_{s\to 0}S\Delta(s) \tag{27}$$

equation 26 becomes

$$\Delta(t)=mK_g\tau_m \tag{28}$$

In Fig. 13 the line $BCDE$ has a slope of mK_g and is tangent to the middle portion of the motor response curve. Since OA is parallel to $BCDE$, it can be shown that

$$AC=\Delta(t)$$

and

$$OB=\tau_m \tag{29}$$

Therefore the time constant of a two-phase motor can be obtained experimentally by applying a step input to the field and recording the output by an oscillograph. A tangent line is next drawn to the straight portion of the motor response curve and is extended to intersect with the time axis. The motor time constant is the distance between the application of step and this intersection.

This method can also be generalized to determine the coefficient of the first order term in any system of the form

$$G(s)=\frac{A(s)}{B(s)}=\frac{\Pi_{k=1}^{m}(T_k s+1)}{\Pi_{j=1}^{N}(T_j s+1)s^n}\quad N\geq m \tag{30}$$

References

1. On Some Lonlinear Phenomena in Regulatory Systems, L. C. Goldfarb. *Avtomatika i Telemechanika*, Moscow, USSR vol. 8, No. 5, 1947, pp. 349–383. Translated in *N. B. S. Report 1691*, National Bureau of Standards, Washington, D. C., 1952.

2. A Frequency Response Method for Analyzing and Synthesizing Contactor Servomechanisms, Ralph J. Kochenburger. *Ph.D. Thesis*, 1949, Massachusetts Institute of Technology; also *AIEE Transactions*, vol. 69, pt. I, 1950, pp. 270–84.

3. Sinusoidal Analysis of Feedback-Controls Systems Containing Nonlinear Elements, E. Calvin Johnson. *AIEE Transactions*, vol. 71, pt. II, July 1952, pp. 169–81.

4. Describing Function Method of Servomechanism Analysis Applied to Most Commonly Encountered Nonlinearities, H. D. Groff. *Ibid.*, vol. 72, pt. II, Sept. 1953, pp. 243–48.

5. Electrostatic Oscillation in Sensitive Contact Mechanisms, T. E. Stuelpnagel. *Ibid.*, vol. 71, pt. II, 1952 (Jan. 1953 section), pp. 397–401.

6. Principles of Servomechanisms (book), G. S. Brown, D. P. Campbell. John Wiley & Sons Inc., New York, N. Y., 1945.

7. Servomechanism Fundamentals (book), Henri Lauer, Robert Lesnick, Leslie E. Matson. McGraw-Hill Book Company, Inc., New York, N. Y., 1947.

Discussion

Robert W. Bass (Princeton University, Princeton, N. J.): Mr. Chao makes several interesting theoretical and experimental points. In my opinion, however, the observed 7% accuracy is in no way a measure of the efficacy of the describing function method, but is rather a reflection of the standard idealization involved in equation 8. In certain cases quite similar to those under discussion I have proved rigorously that any periodic solution of the underlying nonlinear differential equation which is indicated by the describing function method must in fact exist.

Before proceeding, it is necessary to correct the author's remark about Johnson's rigorous derivation and estimate of the validity and the inaccuracies of this method. Indeed, Johnson's discussion proved no more than did those of Goldfarb or Kochenburger; his estimates are meaningless in the absence of knowledge concerning the radius of convergence of one of his power series. Johnson's real accomplishment was to produce a new heuristic motivation for the method: one seeks for the linearized differential equation a linear equation possessing a "generating solution" whose first frequency correction vanishes when it is perturbed in the "direction" of the given nonlinear equation. This supplements Goldfarb's appeal to the Kryloff-Bogoliuboff principle of harmonic balance, and also Klotter's appeal[1] to the method of Ritz-Galerkin. However, a complete proof requires much deeper investigation.

In a theoretical report[2] of limited circulation, I studied a servo schematically identical with Mr. Chao's, except that the relay circuit was assumed to have a single constant time delay rather than the two time constants of equation 15; in other words, I replaced the G_{RC} of equation 15 by $G_{RC}' = \exp(-st_d)$. But if ω_n is large and $t_d = 2\rho/\omega_n$ is written, there is little difference between G_{RC} and G_{RC}'.

On replacing $e\epsilon$ by ϵ, the corresponding dimensionless difference-differential equation is given[2] by writing $k=0$ in

$$\ddot{\epsilon}(t) + \dot{\epsilon}(t) + F(\Sigma(t-t_d)) = 0 \tag{31}$$

$$\Sigma(t) = \epsilon(t) + k\dot{\epsilon}(t) \tag{32}$$

where $[F\Sigma(t)] = 0$ when $-\hat{\epsilon} + (\operatorname{sgn} \dot{\Sigma}(t))\Delta\hat{\epsilon} < \Sigma(t) < \hat{\epsilon} + (\operatorname{sgn} \dot{\Sigma}(t))\Delta\hat{\epsilon}$; and $[F\Sigma(t)] = \operatorname{sgn} \Sigma(t)$ otherwise. Here I use a dimensionless dead-zone $\hat{\epsilon}$ and hysteresis $\Delta\hat{\epsilon}$ which are proportional respectively to the author's $\Delta_C/2$ and Δ_h, and which satisfy $0 \leq 2\hat{\epsilon} \leq 1$, $0 \leq 2\Delta\hat{\epsilon} \leq 1$; similarly, $0 \leq t_d \leq 1$, $0 \leq k < \infty$.

In reference 2 it is proved that the system of equations 31 and 32 encompasses three mutually exclusive possibilities: it may:

1. Behave like an underdamped linear oscillator, $e(t)$ coming to rest ($\dot{\epsilon} = \dot{e} = 0$) with $0 \leq |e(t)| \leq 2\hat{\epsilon}$ after a finite number of oscillations.

2. Possess a single stable limit cycle, i.e., a "soft" self-sustained oscillation.

3. Possess one unstable and one (larger) stable limit cycle, i.e., a "hard" oscillation.

Let

$$\sigma_1 = e^{t_d} - 1 + t_d - 2\hat{\epsilon}, \quad \sigma_2 = t_d e^{t_d} + 1 + t_d - 2\hat{\epsilon},$$

$$\sigma_3 = \sqrt{\sigma_1(\sigma_1 + 2)}, \quad \sigma_4 = (\sigma_3 - \sigma_1)/(\sigma_1 + \sigma_2),$$

$$\delta = \sigma_3 + \log \sigma_4, \quad \sigma = (e^{t_d} - 1 - t_d) - 2\Delta\hat{\epsilon}$$

$$\Delta\hat{\epsilon}^*(u) = \frac{1}{2}[|\log(1-2u)| - 2u],$$

$$t_d^* = t_d + (e^{t_d} - 1)/(2e^{t_d} - 1).$$

Then there are two main cases. 1. Suppose first that $t_d < 2\hat{\epsilon} < t_d^*$. Then if $\Delta\hat{\epsilon} < \Delta\hat{\epsilon}^*(\hat{\epsilon} - t_d/2)$, there are no oscillations when $\delta < \sigma$, but hard oscillations when $0 < \sigma < \delta$. On the other hand, if $\Delta\hat{\epsilon} > \Delta\hat{\epsilon}^*(\hat{\epsilon} - t_d/2)$ there are always soft oscillations. 2. In the second case, when $t_d^* < 2\hat{\epsilon} \leq 1$, there are either no oscillations, or soft oscillations, according to whether $\Delta\hat{\epsilon}$ is less than or greater than $\Delta\hat{\epsilon}^*(\hat{\epsilon} - t_d/2)$.

When this analysis shows that case 2 or case 3 will occur, one might resort to compensation by error-rate feedback, i.e., let $k > 0$ in equation 32, or employ a phase lead resistance-reactance compensating network whose input closely simulates equation 32. For comparison, consider also the scheme in reference 3

$$\Sigma_1(t) = e(t) + k|\dot{e}(t)|\dot{e}(t) \tag{33}$$

To the three mutually exclusive possibilities already listed, a fourth must now be added:

4. The system behaves, roughly, like an overdamped linear oscillator, i.e., after a few swings $e(t)$ "chatters" along the line $e + k\dot{e} = 0$ in the $(e, \dot{e})$ plane until it rests, as in case 1.

Next, let

$$S = \{1 - \exp[-2(\hat{\epsilon} + \Delta\hat{\epsilon})]\}^{-1}$$

$$k_{\min} = (1 - S\{2\hat{\epsilon} - [\exp(-t_d) - 1 + t_d]\}) \times \{1 - S[1 - \exp(-t_d)]\}^{-1}$$

Let $k_{\max}$ be given by solving

$$k_{\max} + \log(\exp(-t_d) + t_d + 2\Delta\hat{\epsilon} - \exp(-t_d)k_{\max}) = 0$$

and let $k_t = \min(1 - 2\epsilon, k_{\max})$

Then when $k_{\min} \leq k \leq k_t$, case 1 occurs, but when $k_{\max} < k < \infty$, case 4 occurs.

Similarly, replacing equation 32 by equation 33, and letting $t_d = 0$, gives[2]

$$S(1 - S(2\hat{\epsilon})) \leq k \leq 1 - 2\hat{\epsilon}$$

as a sufficient condition for case 1 to occur.

These results, complete and exhaustive for equations 31 and 32, were obtained by a piecewise linear analysis, extending and/or correcting that of Hazen,[4] MacColl,[6] Weiss,[6] and Feldbaum.[7] Similar equations for ramp and sinusoidal inputs[8] were found in reference 3. They show just how smooth and slowly varying such dynamic inputs must be for a given relay servo to be able to cope with them.

The theoretically exact picture just summarized is qualitatively almost identical with that found by a describing function analysis. I recommend the following mnemonic: 1. Insert the complex solution $e(t) = ae^{j\omega t}$ into the nonlinear equation to find the "instantaneous characteristic polynomial." 2. Replace complex numbers in the arguments of nonlinear functions by their real parts, and weight the nonlinear terms of this polynomial by doubling each one. 3. Divide the polynomial by $ae^{j\omega t}$. 4. Multiply the polynomial by $d(\omega t)$ and average over a period (i.e., integrate from 0 to 2π and divide by 2π).

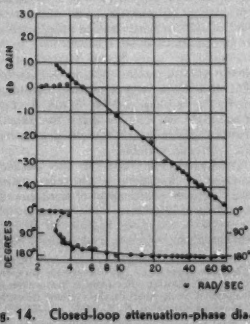


Fig. 14. Closed-loop attenuation-phase diagram

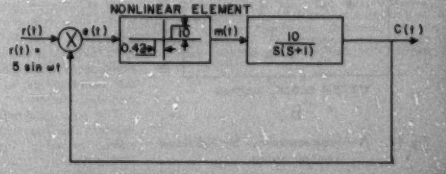
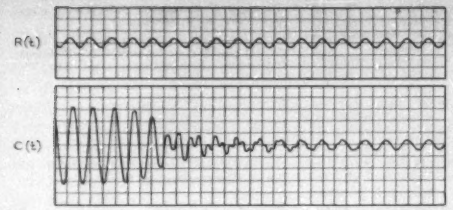
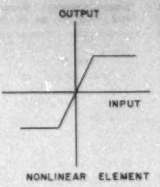
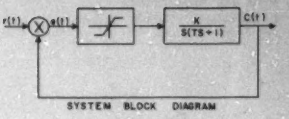


Fig. 15. Nonlinear control system for Fig. 14



A





B

Fig. 16. A—Jump resonance-amplitude decreasing; B—Nonlinear control system

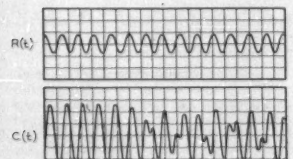
5. Separate this averaged polynomial into real and imaginary parts and find the roots α, ω of the resultant pair of real equations. Alternatively, step 5(A) find these roots graphically by plotting the amplitude and frequency variant parts of the complex polynomial separately as in equation 5.

For difference-differential equations one saves a great deal of manipulation of integrals by replacing the time-delays in the nonlinear terms by time-leads in the linear terms; e.g., replace equation 31 by

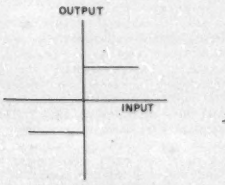
$$\ddot{e}(t+t_d)+\dot{e}(t+t_d)+F[\Sigma(t)]=0 \qquad (31A)$$

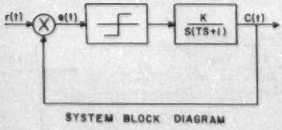
In the case $k=0$ the steps 1 through 5 lead via

$$\omega^2 e^{j\omega t_d}-j\omega e^{j\omega t_d}-2\alpha^{-1}\epsilon^{-j\omega t}F(\alpha\cos\omega t,\ -\omega\alpha\sin\omega t)=0$$



A





B

Fig. 17. A—Jump resonance; B—Nonlinear control system

to $z_\omega = z_\alpha$

where $z_\omega = e^{j\omega t_d}(\omega^2 - j\omega)$,

$$z_\alpha = [a(\alpha) - jb(\alpha)]/\alpha$$

$$\begin{bmatrix} a(\alpha) \\ b(\alpha) \end{bmatrix} = \frac{1}{\pi}\int_0^{2\pi} F(\alpha\cos u,\ -\omega\alpha\sin u)\begin{pmatrix}\cos u \\ \sin u\end{pmatrix} du$$

In the present instance

$$z_\alpha = 2\times \left[\frac{\sqrt{\alpha^2-(\bar{e}+\Delta\bar{e})^2}+\sqrt{\alpha^2-(\bar{e}-\Delta\bar{e})^2}-2j\Delta\bar{e}}{\pi\alpha^2}\right]$$

On plotting z_α and z_ω, one obtains a figure quite similar to the mirror-image of Fig. 8 turned upside down.

The author concludes by suggesting that coulomb friction be considered. If one adds a term C sgn $\dot{e}(t+t_d)$ to equation 31(A), one finds, in reference 2, page 229, that

$$\text{“}z_\omega\text{”} = z_{\omega\alpha} = e^{j\omega t_d}[\omega^2 - j(\omega + C/\pi\alpha)]$$

$$z_\alpha = \{a(\alpha) + j[4C/\pi - b(\alpha)]\}/\alpha$$

Hence, instead of merely two loci, one must plot a whole family of loci in the manner of Loeb.[9] In reference 2 it was shown that as C increases the limit cycles are diminished in amplitude or eliminated. But $C>0$ obviously degrades the transient response.

When linear compensating networks are added, one need only multiply z_ω by the corresponding transfer function. For example, when $k>0$, z_ω becomes

$$z_\omega' = \frac{[\omega^2(1-k) - j\omega(1+k\omega^2)]}{(1+k^2\omega^2)}$$

A study of the equations z_ω' and z_α shows that the describing function method gives an entirely correct qualitative picture, except that it fails to distinguish between the (desirable) case 1 and the (undesirable) case 4.

Turning now to the quantitative side of the picture, consider the special case $k=0$, $t_d=0$, $\Delta\bar{e}=\bar{e}$. This differs from the author's case $\omega_n \to \infty$ only in that $\Delta_b/\Delta_C = \Delta e/2\bar{e} = 0.50$, whereas, from equation 18 $\Delta_b/\Delta_C = 0.16$. By using the Leray-Schauder theory, I proved[2] that for all $\bar{e}$ in $0 \leq 2\bar{e} \leq 1$, if z_ω and z_α intersect just once and are not tangent there, then the original differential equation has at least one nontrivial periodic solution.

The same technique[10] can be used to validate this approximate method in other cases, probably including that under discussion.

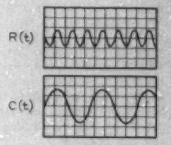


Fig. 18. One-third order subharmonic oscillation

The describing function method has, however, grave drawbacks. Case 4 is undesirable (since, e.g., it wears out the relay) and yet a contactor servo designed as Kochenburger originally proposed may well be found to perform as in case 4.

But by considering the underlying differential equations, and applying a very general algorithm of the writer[11] (a partial version also appeared in reference 2) it can be shown that case 1 can always be attained when feedback from the relay itself is combined with the conventional error and error-rate feedbacks in certain proportions.

In fact, replace equation 32 by

$$\Sigma_s(t) = e(t) + k\dot{e}(t) - k_1 F[\Sigma_s(t-t_d)] \qquad (34)$$

where $k = [1-\exp(-t_d)](>0)$, $k_1' = [\exp(-t_d)-1+t_d]$ (>0), and $k_1 = k_1' - 2\Delta\bar{e}$. Then, in accordance with calculations by the writer, not yet published, $\Sigma_s(t) = e(t+t_d) + 2\Delta\bar{e}F$, so that $\Sigma_s(t-t_d) = e(t) + 2\Delta\bar{e}F$. Moreover, equations 31 and 34 now behave exactly like equations 31 and 32 with $k = t_d = \Delta\bar{e} = 0$. This, then, is theoretically the ideal method of stabilizing a contactor servo.

Since this method completely eliminates the effects of system delays, a nonlinear optimizing network in the manner of McDonald[12] or Stout[13] can be added without using dual-mode control.

Finally, for ramp inputs ($\dot{r}(t) = \text{const}$) or slowly varying inputs $r(t)$, it was shown[2] that the delays can again be eliminated merely by adding a term $k_1'\dot{r}(t)$ to $\Sigma_s(t)$ of equation 34, i.e., by using the rate of change of the reference variable as a secondary reference variable.

Particularly in view of the author's remarks in Appendix I, an experimental study of the practicality of equation 34 should be of interest.

References

1. Steady State Vibrations in Systems Having Arbitrary Restoring and Arbitrary Damping Forces. **K. Klotter.** *Proceedings*, Symposium on Nonlinear Circuit Analysis, Polytechnic Institute of Brooklyn, Brooklyn, N. Y., 1953, pp. 234–57.

2. The Analysis and Synthesis of Relay and Nonlinear Servomechanisms, **R. W. Bass.** *Final*

Report, section 111, Johns Hopkins Institute for Cooperative Research, Baltimore, Md., *Contract DA-36-034-ORD-1274RD*, June 30, 1955.

3. The Stabilization of On-Off Controlled Servomechanisms, **A. M. Uttley, P. H. Hammond.** *Automatic and Manual Control*, Butterworth's Scientific Publications, London, England, 1952, pp. 285–307.

4. Theory of Servomechanisms, **H. L. Hazen.** *Journal*, Franklin Institute, Philadelphia, Pa., vol. 218, 1934, pp. 279–330.

5. Fundamental Theory of Servomechanisms (book), **L. A. MacColl.** D. Van Nostrand & Company, New York, N. Y., 1945.

6. Analysis of Relay Servomechanisms, **H. K. Weiss.** *Journal*, Aeronautical Sciences, New York, N. Y., vol. 13, 1946, pp. 364–76.

7. The Simplest Relay Systems of Automatic Control, **A. A. Feldbaum.** (in Russian). *Automatika i Telemechanika*, Moscow, USSR, vol. 10, 1949, pp. 249–66.

8. Relay Servomechanisms, **T. A. Rogers, W. C. Hurty.** *Transactions*, American Society of Mechanical Engineers, New York, N. Y., vol. 72, 1950, pp. 1163–72.

9. Recent Advances in Nonlinear Servo Theory, **J. M. Loeb.** *Ibid.*, vol. 76, 1954, pp. 1281–89.

10. Equivalent Linearization, Nonlinear Circuit Synthesis, and the Stabilization and Optimization of Control Systems, **R. W. Bass.** *Proceedings*, Symposium on Nonlinear Circuit Analysis, Polytechnic Institute of Brooklyn, Brooklyn, N. Y., 1956.

11. Improved On-Off Missile Stabilization, **R. W. Bass.** *IAS Preprint No. 589, ARS Preprint No. 283-56*, Institute of Aeronautical Sciences, New York, N. Y., American Rocket Society, Inc., New York, N. Y., 1956.

12. Intentional Nonlinearization of Servomechanisms, **D. C. McDonald.** *Proceedings*, Symposium on Nonlinear Circuit Analysis, Polytechnic Institute of Brooklyn, Brooklyn, N. Y., 1953, pp. 402–11.

13. Effects of Friction in an Optimum Relay Servomechanism, **T. M. Stout.** *AIEE Transactions*, vol. 72, pt. II, Nov. 1953, pp. 329–36.

A. M. Hopkin and **K. Ogata** (University of California, Berkeley, Calif.): We have been examining the response of somewhat similar systems to sinusoidal forcing. Two very noticeable phenomena were jump-resonance and the presence of subharmonic response to sinusoidal forcing. As might be expected for nonlinear systems, the output amplitude and frequency for the same form of input has at times been different for different initial conditions and has changed abruptly in the presence of small disturbances. Examples of these phenomena taken from analogue computer studies are shown in the accompanying figures.

Fig. 14 shows the closed-loop attenuation-phase characteristic for the system whose block diagram is shown in Fig. 15. Although the output of this system is almost sinusoidal in steady state at all times, the curve is double-valued in a region and the output amplitude and phase may jump abruptly at either end of the double-valued region.

Figs. 16(A) and 17(A) show the transient portion of a jump-resonance phenomena for the systems whose block diagrams are shown in Figs. 16(B) and 17(B) respectively. It will be noticed that the steady-state wave forms before and after the amplitude jump are very nearly sinusoidal, but that the transient portion of the jump is not sinusoidal in form.

Table IV. Summary of Results*

Condition	Theoretical Results		Experimental Result		Per cent Deviation	
	f	a	f	a	f	a
a	2.79	14.30	1.95	18.00	+30.0	−26.0
b	2.79	10.70	1.83	13.40	+34.2	−25.2
c	2.79	7.12	1.68	9.55	+39.7	−33.5
d	2.70	3.60	1.80	4.78	+46.0	−32.7
e	2.70	1.96	1.56	3.94	+41.8	−101.0
f	2.31	11.95	1.70	16.20	+26.4	−35.0
g	2.15	11.30	1.57	13.50	+27.0	−19.5
h	1.75	7.80	1.37	8.45	+21.7	−8.2

* For a servo system similar to Fig. 4 but with two contactors in cascade.
f = frequency of oscillation in cycles per second.
a = nondimensional amplitude of oscillation.

Fig. 18 shows the input and output of a system with a subharmonic oscillation of the order one-third. Once again the output is very nearly sinusoidal in form. This type of subharmonic oscillation may be observed in each of the systems shown in Figs. 15, 16(B), and 17(B), providing the inputs have the proper amplitudes, and the proper initial conditions are set up.

Since the jump-resonance phenomenon and the subharmonic modes of oscillation can exist for sinusoidal forcing of systems which seem to fulfill the requirements which make the describing function valid, it is felt that describing function analysis must be used with care, particularly when designing systems which may operate with periodic inputs or disturbances.

Stanley K. Chao: The discussions given by Dr. Bass, Professor Hopkin, and Mr. Ogata are valuable additions to the paper. In his discussion, Dr. Bass gives a detailed comparison between the servo system used in the investigation presented in the paper and the system he employed for his theoretical study and finds them "schematically" identical. It should be pointed out, however, that the assumption of the relay circuit transfer function as simply a time delay is not always valid; particularly when the resonant frequency of the relay circuit ω_n is smaller than the frequency of oscillation of the servo system. Without knowing the exact operating point of the servo system, it is rather risky to make such an assumption. The author has also investigated a servo system using two contactors in cascade. While the exact transfer function and the theoretical describing function were used for the leading contactor, the second contactor was simply assumed to be a time delay. The question of what was the delay time of a relay arose in view of the fact that the pull-in time and the drop-out time of a relay are not the same. Should it be one, or the other, or the average of both? Using exactly the same procedure as that in this paper, the results obtained, corresponding to Table III, are shown in Table IV. The first and leading relay is the same as shown in Fig. 4. The second relay had pull-in time = 16 milliseconds, drop-out time = 9 milliseconds. For the result obtained here the second relay was assumed to be representable by a delay of 16 milliseconds.

While there is no positive proof that these relatively large discrepancies between the theoretical and experimental results are due entirely to the time delay assumption for the second relay, it is felt that more exhaustive study is called for, since the frequency of oscillation of the system was lower (Table IV) than the resonant frequency of the relay circuit (6.8 cycles per second).

Dr. Bass's work represents an important contribution to the theoretical analysis of relay and nonlinear servomechanisms. It is comforting to know that there exists another theoretical study which seems to verify further the validity of the describing function method, which is a powerful tool for nonlinear system studies. However, it is important to realize its limitations. Some of these limitations are pointed out by Dr. Bass in his discussion and in papers by others.[1] They are also clearly demonstrated in the paper.

In applying the describing function method to practical design problems, one should bear in mind that it is inherently an approximate method and that the describing function should completely "describe" all nonlinearities in the system.

Unlike the phase-plane method which deals primarily with the transient condition, the describing function method is concerned principally with the steady-state oscillatory condition. The jump-resonance observed in Professor Hopkin's and Mr. Ogata's discussion illustrates this limitation. Since it is not intended to predict transient condition, it should not be construed to mean the failure of the method. The next natural step in the theory of the describing function method would be some refinement to remedy this limitation, thus extending its usefulness.

Reference

1. Basic Method for Nonlinear Control System Analysis, **T. M. Stout.** AIEE-ASME Nonlinear Controls Conference, Princeton, N. J. Mar. 26–28, 1956. *Transactions*, American Society of Mechanical Engineers, New York, N. Y., (*Paper no. 56.IRD 9*).

The Frequency Response of a Servomechanism Designed for Optimum Transient Response

J. C. WEST
ASSOCIATE MEMBER AIEE

P. N. NIKIFORUK
NONMEMBER AIEE

Synopsis: The frequency response of a mechanical system which is subjected to torque limitation and which, by the addition of nonlinear error processing, is designed to have an optimum response to step-function inputs, is evaluated qualitatively. This analysis is carried out to determine the effect on the frequency response of the error-processing element and, moreover, to determine whether this system has advantages in addition to optimum transient response. The closed-loop response of this system to a sinusoidal excitation signal is evaluated from a graphical representation of the open-loop frequency response. The latter takes into consideration the harmonics in the control signal. This closed-loop response is compared with that of the equivalent system subjected only to torque limitation. Both velocity feedback and derivative of error stabilization are considered. The mechanical system is simulated on an analogue computer and experimental verification of the analysis is obtained. Finally, the approximate frequency response of the optimum system is evaluated by neglecting the harmonics in the control signal. This approximate response is compared with that evaluated previously, and this comparison shows that the neglect of these harmonics is justifiable.

ALL servomechanisms are nonlinear to the extent that saturation of the servoamplifier occurs at large control signal levels. Recent work[1-3] has shown that the transient response of systems operating under such conditions can be made faster and less oscillatory by the deliberate introduction of additional nonlinear elements. In particular, a servomechanism of this kind is one incorporating serme control (one where control signal is proportional to the normal stabilization signal and the sign-error-root-modulus-error signal). The response to step-function inputs of systems incorporating serme control has been studied extensively,[1] and this principle has been applied to a mechanical system in which the output is required to move between any two of 64 positions in a minimum of time. This application[4-5] of serme control to a mechanical system subjected to torque limitation has resulted in a considerable improvement in the transient response of the system. For several reasons it was considered necessary to evaluate the response of this optimum system to a sinusoidal excitation signal. First, it is necessary to show that this system has a good frequency response. This follows from the fact that since the optimum system is nonlinear, a good transient response is not necessarily indicative of a good frequency response. Second, since the optimum system is derived from one subjected to torque limitation it is necessary to show that it will follow sinusoidal inputs in a manner that is as good as that of the latter system. Finally, it was thought that the addition of this processing element might result in an improvement in the frequency response of the nonlinear system. It has been found that this improvement actually occurs.

Techniques[6-8] exist for evaluating the response to sinusoidal excitation signals of control systems incorporating only one nonlinear element. These techniques utilize the describing function,[9] the basis of which is the assumption that for a sinusoidal input to a nonlinear element the fundamental component of the Fourier expansion of the element's output can be taken as representative of the actual response. In the system incorporating serme control, as well as in many other control systems, the input to the second nonlinear element is a highly distorted signal. Since the response of a nonlinear element to the fundamental component of a distorted signal is a function of the harmonics present, the evaluation of the frequency response of control systems incorporating more than one nonlinearity is exceedingly tedious. In many cases a qualitative evaluation is possible only on a digital computer. The solution of entire control system problems by such means is not attractive and it is desirable, therefore, to determine the effect of neglecting the harmonic content in the signal at the input to every nonlinear element.

The purpose of the work described herein is therefore twofold. First, it is of interest to evaluate the response to sinusoidal excitation signals of a servomechanism incorporating serme control to determine whether this system has advantages in addition to optimum transient operation. Second, it is of interest to determine the effect on the predicted frequency response of neglecting the harmonic content of the signal at the input to the second nonlinear element.

Paper 56-452, recommended by the AIEE Feedback Control Systems Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE–ASME Nonlinear Controls Conference, Princeton, N. J., March 26–28, 1956. Manuscript submitted August 3, 1955; made available for printing February 14, 1956.

J. C. WEST and P. N. NIKIFORUK are with The University of Manchester Electrical Engineering Laboratories, Manchester, England.

Nomenclature

θ_i = instantaneous input signal to the servomechanism
θ_0 = instantaneous output signal from the servomechanism
$\theta = \theta_i - \theta_0$ = instantaneous error signal of the servomechanism
θ_c = instantaneous control signal of the servomechanism
X = output signal in phase-plane notation
Y = output velocity in phase-plane notation
T_d = damping time constant
V_c = control signal level at which torque limitation occurs
α, g = gain constants
E_i = amplitude of input signal
E = amplitude of error signal
V_a = amplitude of error signal at which error processing occurs
$f = \omega/2\pi$ = frequency of input signal
E_g = amplitude of the fundamental component of the output from the nonlinear error-processing element
E_c = amplitude of the fundamental component of the error signal
β = incomplete beta function
Γ = gamma function

The Basic System

The basic system to be considered is of second order and is shown in Fig. 1. It is stabilized by velocity feedback and is subjected to torque limitation, the characteristic of the servoamplifier being of the form shown in Fig. 2. When the amplitude of the excitation signal is small and the following condition prevails:

$$|\theta_c| < V_c$$

the response of the system is linear The equation of motion is then

$$\ddot{\theta}_0 = \alpha\omega_0^2\{\theta_i - \theta_0 - T_d\dot{\theta}_0\} \qquad (1)$$

Linear operation ceases when

$$\theta_c > V_c$$

$$\theta_c < -V_c$$

and the equation of motion becomes

$$\ddot{\theta}_0 = \pm\omega_0^2 V_c \qquad (2)$$

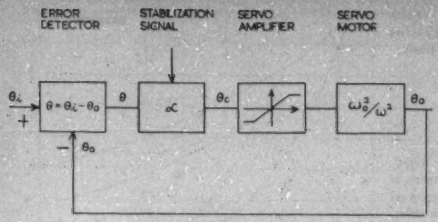


Fig. 1. Basic system

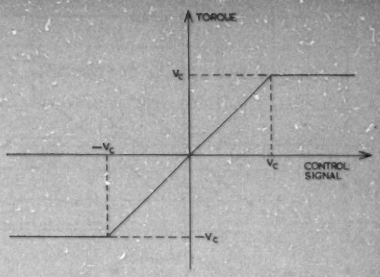


Fig. 2 (right). Saturation characteristic

For convenience in phase-plane notation let

$$X = \theta_i$$

$$Y = \dot{\theta}_i \tag{3}$$

Equation 2 can now be expressed in the form

$$Y\frac{dY}{dX} = \pm\omega_0^2 V_c \tag{4}$$

The solution of this equation is

$$Y^2 - Y_1^2 = \pm 2\omega_0^2 V_c (X - X_1) \tag{5}$$

where X_1 and Y_1 are the initial values of position and velocity.

Utilization of Maximum Torque

Since the optimum transient response is considered to be that in which the output moves between two positions in a minimum of time, it is necessary that the motion of the system be characterized by a period of maximum acceleration followed by a period of maximum deceleration. In the system considered, maximum deceleration without overshooting occurs along that trajectory in the phase plane which terminates at the origin. This trajectory is referred to as the critical boundary and is given, from equation 5, by

$$Y^2 = \pm 2\omega_0^2 V_c X \tag{6}$$

For minimum response time it is necessary, therefore, that maximum acceleration ensue up to this boundary, at which point the system must be switched to maximum deceleration. To obtain such a force changeover, it is necessary to have in the system an additional nonlinear element. The form of this element is such that the control signal is of one sign if the trajectory of motion in the phase plane is below this critical boundary, and is of the opposite sign if the trajectory is above this boundary and is zero on the boundary.

Let the control signal be expressed in the form

$$\theta_c = f(\theta) + g(\dot{\theta})_0$$

This signal must be zero for all points on the critical boundary and consequently it is necessary that

$$f(\theta) = -g(\dot{\theta}_0)$$

be identical with equation 6.

Serme Control

In this type of control the functions f and g are taken to be of the form

$$f(\theta) = g \operatorname{sgn}(\theta_i - \theta_0)\sqrt{|\theta_i - \theta_0|} \tag{7}$$

$$g(\dot{\theta}_0) = T_d\dot{\theta}_0 \tag{8}$$

The control signal is therefore

$$\theta_c = \alpha\{g \operatorname{sgn}(\theta_i - \theta_0)\sqrt{|\theta_i - \theta_0|} - T_d\dot{\theta}_0\} \tag{9}$$

By comparing this expression with equation 6 it is found that for minimum response time

$$g = \omega_0 T_d\sqrt{2V_c} \tag{10}$$

When the error signal is small the gain of the error-processing element is very high. Consequently a mechanical system incorporating this characteristic would become unstable at small error and control signal levels. This is avoided by making the processing element linear at small error signal levels. The characteristic is of the form shown in Fig. 3, linear operation ceasing at the error signal levels $\pm V_a$. The amplitude of the error signal at which this occurs is, from equation 6

$$V_a = 2(\omega_0 T_d)^2 V_c \tag{11}$$

Evaluation of the Frequency Response

In this analysis the closed-loop response to a sinusoidal excitation signal is evaluated from a graphical representation of the open-loop frequency response that is similar to the Nyquist diagram. The basic assumption[9] of this technique is that the harmonics generated by the torque-limiting amplifier are effectively filtered out by the servomotor. This enables the error signal to be represented by a single sinusoid. For the purpose of analysis the closed loop of the nonlinear system is broken at a point immediately following the error detector and a sinusoidal signal is applied to the processing element. The fundamental component of the system output and its phase with respect to the error signal are evaluated either graphically or by means of a digital computer, the effect of the harmonics in the control signal being taken into consideration. The amplitude and phase of the excitation signal are then evaluated from the vector relationship

$$\boldsymbol{\theta}_i = \boldsymbol{\theta} + \boldsymbol{\theta}_0 \tag{12}$$

In this equation $\boldsymbol{\theta}_i$ is a function of the amplitude (E) and the frequency (ω) of the error signal and is taken to be the representation of the open-loop frequency response.

The reference vector ($\boldsymbol{\theta}$) is taken to be along the axis of the reals. For an excitation signal of constant frequency the resulting plot of $\boldsymbol{\theta}_i$ in the complex plane is a vector locus of the type shown by full lines in Fig. 4. For an excitation signal of constant amplitude (E_i) the tip of $\boldsymbol{\theta}_i$ must lie on a circle of radius E_i, the center being the origin. Points of intersection of the appropriate circle and constant frequency loci give the operating points of the system, multiple intersections indicating transitions in the gain-frequency characteristic.

Consider an excitation signal whose amplitude is kept constant and where frequency is increased slowly from zero. Since points of intersection of the appropriate circle and constant frequency loci give the operating points of the system, it is apparent from Fig. 4 that the operating point moves in a clockwise direction along the circle loci with increasing signal fre-

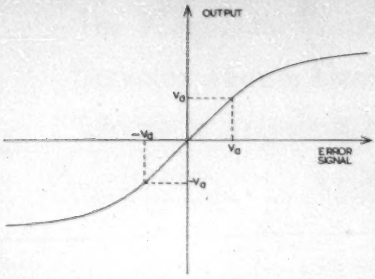


Fig. 3 (left). Error-processing characteristic

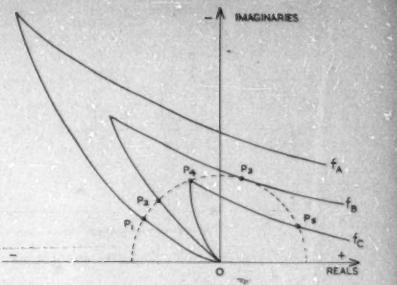


Fig. 4 (right). Representation of the open-loop response for different input frequencies, f

$f_A > f_B > f_A$

quency. At very low frequencies the appropriate loci intersect at only one point and consequently there is only one possible operating point. Thus, for example, at the frequency f_A the appropriate loci intersect at only the one point P_1 and this is the operating point of the system. As the frequency of the excitation signal is increased further the operating point continues to move in a clockwise direction, until at the frequency f_B a condition is reached at which there are two intersections, P_2 and P_3, the latter point being a tangency. There is, however, no ambiguity of operation for the operating conditions corresponding to P_3 are unstable for increasing signal frequency. Consequently as the signal frequency is increased, the operating point continues to move uniformly along the circle locus until the frequency f_C is attained. At this frequency the operating point is P_4 and it is a tangency. If the frequency of the excitation signal is increased very slightly there is no intersection in the region of P_4 and the operating point must move discontinuously to the region of P_5. There is, therefore, a transition in the gain-frequency characteristic. Following this transition the response is uniform for increasing signal frequency, the operating point continuing to move along the circle locus in a clockwise direction. The method of evaluating the response for decreasing signal frequency is similar to that just described for increasing signal frequency. If the frequency of the excitation signal is decreased from a value just slightly greater than f_C to the value f_B, the operating point moves along the circle locus in a counterclockwise direction from P_4 to P_3. The point P_3 is a tangency and if the signal frequency is decreased very slightly the operating point moves discontinuously from P_3 to the region of P_2. This second transition completes the hysteresis loop in the gain-frequency characteristic and the subsequent motion of the operating point for decreasing signal frequency is in a counterclockwise direction along the circle locus.

The frequency response of the system incorporating serme control has been evaluated in this manner and a typical set of gain-frequency characteristics is shown in Fig. 5. For the particular value of damping considered there is no jump phenomenon in the frequency response. The responses shown are, nevertheless, typical of systems incorporating amplitude-limiting characteristics, the bandwidth of the response decreasing with increasing signal amplitude.

Frequency Response of the Equivalent System Subjected Only to Torque Limitation

The equivalent system incorporating only torque limitation is one in which the nonlinear error-processing element is replaced by a linear amplifier. The gain of this amplifier is equal to that of the linear regime of the error-processing element. The response of this system has been evaluated by use of the technique described on the preceding section and a typical set of gain-frequency characteristics is shown in Fig. 6. By comparing these responses with those shown in Fig. 5, it is observed that jump phenomenon is present in the gain-frequency characteristic of the system subjected to torque limitation for values of damping at which such phenomenon does not occur in the response of the system incorporating serme control Figs. 5 and 6 also show that the reduction in the bandwidths of the response is the same for both systems.

Derivative of Error Stabilization

In some cases,[10] velocity feedback stabilization is undesirable because of the velocity lag that occurs when the system input is moving at a uniform rate. An alternative method of stabilization is that of derivative of error. This is obtained by inserting into the closed-loop system a differentiating network in parallel with the nonlinear error processing element. For such a system the equation of motion in the linear regime is

$$\ddot{\theta}_0 = \omega_0\omega_0^2[(\theta_i-\theta_0)+T_d(\dot{\theta}_i-\dot{\theta}_0)] \qquad (13)$$

where T_d is the damping time constant of the derivative of error stabilizing network.

The equations of motion in the nonlinear regime of operation are

$$\ddot{\theta}_0 = \pm\omega_0^2 V_c$$

and from an analysis similar to that given in the section on "Serme Control" it is found that for optimum transient response

$$V_0 = 2(\omega_0 T_d)^2 V_c \qquad (14)$$

Frequency Response of the Optimum System

The response of the optimum system has been evaluated using the technique described in the section on "The Evaluation of the Frequency Response" and a typical set of gain-frequency characteristics is shown in Fig. 7. This figure shows that for the value of damping considered there is jump phenomenon in the frequency characteristic. In addition, it shows that the bandwidth of the response decreases with increasing signal amplitude in the same manner as it does for velocity feedback stabilization.

Frequency Response of the System Subjected only to Torque Limitation

The response of this system has been evaluated and typical gain-frequency characteristics are shown in Fig. 8. For the value of damping considered jump

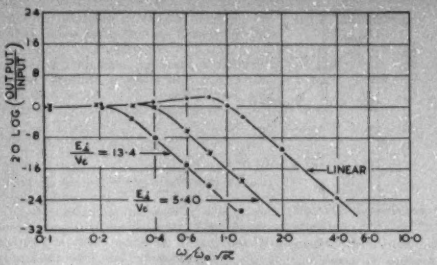


Fig. 5. Response of the optimum system stabilized by velocity feedback. One-half critical damping

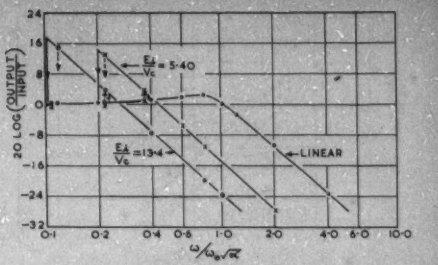


Fig. 6. Response of the system subjected only to torque limitation and using velocity feedback stabilization. One-half critical damping

Experimental points: •—Linear x—Nonlinear, $E_i/V_c = 5.40$ 0—Nonlinear, $E_i/V_c = 13.4$ Solid line indicates theoretical responses
(Subcaption applies to both figures)

phenomenon occurs in the frequency response. By comparing the responses shown in this figure with those shown in Fig. 7, it is observed that the hysteresis loop in the gain-frequency characteristic of the system incorporating serme control is much narrower than that in the corresponding characteristic of the system subjected only to torque limitation. The upper transitional frequencies are approximately the same for both systems. However, the lower transitional frequency for the optimum system is considerably greater than that for the system subjected to torque limitation. The decrease in bandwidths is the same for both systems.

Approximate Frequency Response of the Optimum System

In the optimum system the error signal, i.e., the input to the nonlinear error processing element, is sinusoidal. The input to the torque limiting amplifier is, however, a highly distorted signal, the harmonic components of which modify the response of this element to the fundamental component of the input. The preceding analysis takes these harmonics into consideration and is exceedingly laborious. If it could be shown that it is justifiable, to a first approximation, to neglect the harmonic components of the control signal, then the preceding analysis would become much less laborious and the frequency response could be evaluated without the aid of a digital computer. This extension of the describing function technique could then be applied to evaluate the approximate frequency response of any servomechanism incorporating more than one nonlinearity. The results of the following analysis show that this is possible.

Approximate Response with Velocity Feedback Stabilization

The system stabilized by velocity feedback consists essentially of two sections. The first section includes only the nonlinear error-processing element. It is frequency invariant and has the transfer function $D(E)$, where $D(E)$ is the describing function of the error processing element. It is derived in the Appendix and is given by

$$D(E) = \frac{2}{\pi}\left\{\sin^{-1}\left(\frac{V_a}{E} - \frac{V_a}{E}\sqrt{1-\left(\frac{V_a}{E}\right)^2}\right) + \sqrt{\frac{\pi V_a}{E}}\,\frac{\Gamma(5/4)}{\Gamma(7/4)} - \sqrt{\frac{V_a}{E}}\,\beta_2(3/4,\,3/4)\right\} \quad (15)$$

The second section includes the remaining elements of the system and is frequency and amplitude dependent. Neglecting harmonics, it has the approximate transfer function $H(E_g, \omega)$ where E_g is the amplitude of the fundamental component of the output from the nonlinear error-processing element. The transfer function $H(E_g, \omega)$ is determined, from Fig. 9 in the following manner. For a specified error signal, the response of the nonlinear processing element is known, i.e., E_g and ω are known. Knowing these two variables, it is possible to evaluate the response of the inner loop because the signals at the input and the output of the torque-limiting amplifier must satisfy the amplitude relationship imposed by the amplifier and the frequency relationship imposed by the remaining elements of the inner stabilization loop. Let it be assumed that the fundamental component at the output from the torque-limiting amplifier is in phase with the fundamental component of the control signal. Using the notation indicated in Fig. 9, the frequency relationship imposed by the inner loop is found to be

$$E_c e^{j\omega t} = \alpha\left\{E_g e^{j} + jk\frac{\omega_0}{\omega}B\right\}e^{j\omega t}$$

Since E_c is real

$$\frac{E_c}{E_g} = \alpha\sqrt{1-\left[\frac{KB}{E_g}\,\frac{\omega_0}{\omega}\right]^2} \quad (16)$$

Neglecting harmonics, the amplitude relationship that is imposed by the saturation characteristic is determined from a Fourier analysis of its output and is found to be

$$\frac{B}{E_g} = \frac{2}{\pi}\left\{\frac{E_c}{E_g}\sin^{-1}\left(\frac{V_c}{E_c}\right) + \frac{1}{E_g}\sqrt{1-\left(\frac{V_c}{E_c}\right)^2}\right\} \quad (17)$$

The approximate operating point of the inner loop is determined by superimposing the loci of equation 16 onto the loci of equation 17.

The open-loop response is then

$$\theta_0 = D(E)H(E_g, \omega)\theta$$

and the open-loop representation of this response is, therefore,

$$\theta_i = [1 + D(E)H(E),\omega)]\theta \quad (18)$$

The approximate open-loop frequency response has been evaluated in this manner and typical loci are shown by dashed line in Fig. 10. The loci evaluated previously are also shown in this figure but in full line. There is little difference between the two sets of loci and, consequently, the closed-loop responses are nearly identical.

Approximate Response with Derivative of Error Stabilization

Let the closed-loop system be broken at a point immediately following the error detector and let a signal $Ee^{j\omega t}$ be applied to the error-processing elements. The fundamental component in the output of

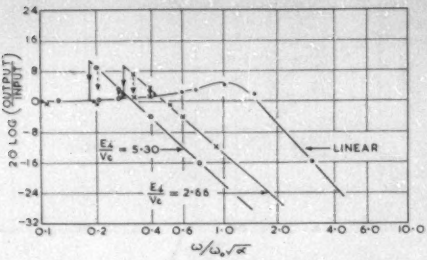


Fig. 7. Response of the optimum system stabilized by derivative of error. Two-fifths critical damping

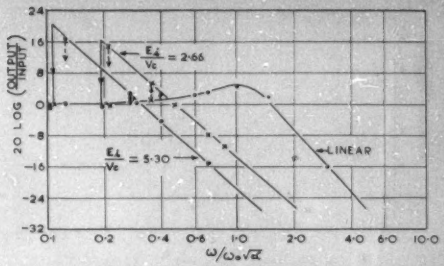


Fig. 8. Response of the system subjected only to torque limitation and using derivative of error stabilization. Two-fifths critical damping

Experimental points: •—Linear x—Nonlinear, $E_1/V_c = 2.66$ 0—Nonlinear, $E_1/V_c = 5.30$ Solid line indicates theoretical responses
(Subcaption applies to both figures)

the nonlinear error-processing element is

$$D(E)E^{j\omega t}$$

and the response of the stabilizing network is

$$\omega T_d E^{j\omega t}$$

The fundamental component of the control signal is therefore

$$\alpha[D(E)+j\omega T_d]E^{j\omega t}$$

Neglecting harmonics, the transfer function of the torque-limiting amplifier becomes the describing function $D(E_c)$, which is expressed by equation 17. The fundamental component of the system output is

$$-\left(\frac{\omega_0}{\omega}\right)^2 D(E_c)\alpha[D(E)+j\omega T_d]E^{j\omega t}$$

and the representation of the open-loop response is therefore

$$\theta_i = \left\{1-\left(\frac{\omega_0}{\omega}\right)^2 D(E_c)\alpha[D(E)+j\omega T_{d1}]\right\}E^{j\omega t} \qquad (19)$$

The approximate response has been evaluated in this manner and typical open-loop loci are shown by dashed line in Fig. 11. These loci correspond approximately with those evaluated previously, the latter being shown by full line. The closed-loop responses are consequently nearly identical.

Experimental System

The mechanical system has been simulated on an analogue computer, the basic element of which is the anode follower. The nonlinear error-processing element is based on a design given in reference 11 and is essentially an anode follower of variable gain. Diode switches are inserted in the ratio arms so that the gain of the element decreases as the amplitude of the input signal increases. The design of the simulator is conventional and consequently no further detail is given.

Experimental Results

The response of the optimum system to a sinusoidal excitation signal has been determined on the analogue computer and typical results are shown in Figs. 5 and 7 together with the calculated loci. The corresponding responses of the system subjected to torque limitation have been determined also on the analogue computer and are shown in Figs. 6 and 8 together with the calculated loci. For both systems and for both types of stabilization there is good agreement between the experimental and the predicted responses.

Conclusions

The frequency response of a mechanical system incorporating serme control has been evaluated by a technique which takes into consideration the harmonics in the control signal, and which is based on the assumption that the error signal is sinusoidal. The system has been simulated on an analogue computer and experimental frequency responses have been obtained. There is good agreement between the experimental and the calculated responses and it is concluded from this that the technique is reliable.

The preceding analysis has also shown that a system incorporating serme control has advantages in addition to optimum transient response. It has shown that the optimum system responds to sinusoidal inputs in a manner that is superior to that in which responds the system subjected only to torque limitation. The frequency response of the optimum system has a bandwidth as great as, if not greater than, that of the system subjected only to torque limitation. Moreover, the nonlinear error-processing element either quenches the jump phenomenon that occurs in frequency response of the latter system, or reduces it to a magnitude that is much smaller. It is concluded, therefore, that the addition of serme control improves the response of a system-possessing torque limitation to sinusoidal as well as to step-function inputs.

Finally, the effect of neglecting the harmonics in the control signal of the optimum system has been evaluated. It has been shown that the errors introduced by neglecting these harmonics are small and that this approximation is justifiable. It is believed that this extension of the describing function technique is applicable to other systems in which the nonlinearities are separated by elements having multiple inputs.

Appendix I. Describing Function for the Nonlinear Error-Processing Element

Let the input to the nonlinear element shown in Fig. 2 be

$$x = E \sin \omega t$$

and let the nonlinear characteristic be

$$Y = f(x)$$

The output is then

$$Y = f(E \sin \omega t)$$

This is a periodic wave form and can be expressed by the series

$$Y = a_0 + a_1 \cos \omega t + a_2 \cos 2\omega t + \ldots\ldots + b_1 \sin \omega t + b_2 \sin 2\omega t + \ldots\ldots$$

Since $f(x)$ is a single-valued odd function of x and since $\cos n\omega t$ is an even function of (t), $a_0, a_1, a_2, \ldots$ are all zero.

The fundamental component of the output is, therefore, $b_1 \sin \omega t$ and is in phase with the input. The amplitude of the fundamental component is given by

$$\frac{4}{\pi}\int_0^{\pi/2} f(E \sin \omega t) \sin \omega t \, d\omega t$$

Let E_g denote the amplitude of the fundamental component of the output (in place of b_1) and let ϕ denote the angle at which nonlinear operation first occurs. The discontinuity in the nonlinear characteristic occurs at $x = V_a$, and therefore

$$\phi = \text{arc sin } V_a/E \text{ for } 0 < \phi < \pi/2$$

The integral becomes then

$$E_g = \frac{4}{\pi}\int_0^{\phi} E \sin^2 \omega t \, d\omega t + \frac{4}{\pi}\int_{\phi}^{\pi/2} g\sqrt{E} \sin^{3/2} \omega t \, d\omega t$$

$$= \frac{4}{\pi}\int_0^{\phi} E \sin^2 \omega t \, d\omega t + \frac{4}{\pi}\int_0^{\pi/2} g\sqrt{E} \sin^{3/2} \omega t \, d\omega t - \frac{4}{\pi}\int_0^{\phi} g\sqrt{E} \sin^{3/2} \omega t \, d\omega t$$

But $g = \sqrt{V_a}$ (see Fig. 2) and

$$\int_0^{\pi/2} \sin^n x \, dx = \frac{\sqrt{\pi}}{2}\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma(n/2+1)}$$

Hence

$$E_g = \frac{2E}{\pi}\left\{\sin^{-1}\left(\frac{V_a}{E}\right) - \frac{V_a}{E}\sqrt{1-\left(\frac{V_a}{E}\right)^2} + \sqrt{\frac{\pi V_a}{E}}\frac{\Gamma(5/4)}{\Gamma(7/4)} - 2\sqrt{\frac{V_a}{E}} \times \int_0^{\phi} \sin^{3/2} \omega t \, d\omega t\right\}$$

This expression can be simplified by taking Z as the variable of integration where

$$Z = \sin^2 \omega t$$

Hence

$$E_g = \frac{2E}{\pi}\left\{\sin^{-1}\left(\frac{V_a}{E}\right) - \frac{V_a}{E}\sqrt{1-\left(\frac{V_a}{E}\right)^2} + \sqrt{\frac{\pi V_a}{E}}\frac{\Gamma(5/4)}{\Gamma(7/4)} - \sqrt{\frac{V_a}{E}} \times \int_0^{\sin^2 \phi} z^{1/4}(1-z)^{-1/2} dz\right\}$$

But

$$\beta_z(p,q) = \int_0^z x^{p-1}(1-x)^{q-1} dx$$

is the incomplete beta function and is tabulated.

Fig. 9. Stabilization loop

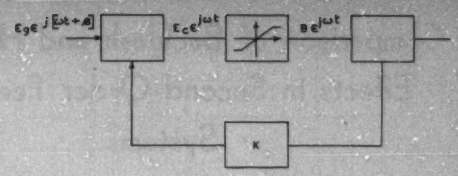


Fig. 10. Comparison of the open-loop loci of the optimum system stabilized by velocity feedback

Solid line indicates effect of taking into consideration the harmonics in the control signal. Dashed line indicates effect of neglecting the harmonics in the control signal
(Subcaption applies to Fig. 11 also)

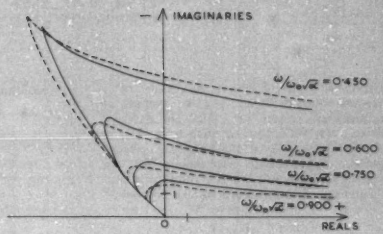
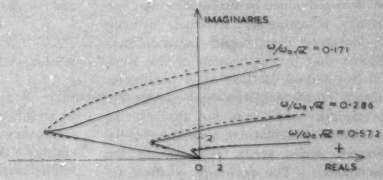


Fig. 11. Comparison of the open-loop loci for the optimum system stabilized by derivative of error



Hence

$$E_g = \frac{2E}{\pi}\left\{\sin^{-1}\left(\frac{V_a}{E}\right) - \frac{V_a}{E}\sqrt{1-\left(\frac{V_a}{E}\right)^2} + \sqrt{\frac{\pi V_a}{E}}\frac{\Gamma(5/4)}{\Gamma(7/4)} - \sqrt{\frac{V_a}{E}}\,\beta_z(5/4,\ 2/4)\right\}$$

The describing function, $D(E)$, is defined as the gain of the sin ωt term, i.e., E_g/E, and is therefore

$$D(E) = \frac{2}{\pi}\left\{\sin^{-1}\left(\frac{V_a}{E}\right) - \frac{V_a}{E}\sqrt{1-\left(\frac{V_a}{E}\right)^2} + \sqrt{\frac{\pi V_a}{E}}\frac{\Gamma(5/4)}{\Gamma(7/4)} - \sqrt{\frac{V_a}{E}}\,\beta(5/4,\ 2/4)\right\}$$

References

1. The Effects of the Addition of Some Non-Linear Elements on the Transient Performance of a Simple R.P.C. System Possessing Torque Limitation, J. C. West, J. L. Douce, R. Naylor. *Proceedings*, Institution of Electrical Engineers, London, England, vol. 101, pt. II, no. 80, 1954, pp. 156–65.

2. Non-Linear Techniques for Improving Servo Performance, D. McDonald. *Proceedings*, National Electronics Conference, Chicago, Ill., vol. 6, 1950, pp. 400–21.

3. A Phase-Plane Approach to the Compensation of Saturating Servomechanisms, Arthur M. Hopkin. *AIEE Transactions*, vol. 70, pt. I, 1951, pp. 631–39.

4. The Use of Non-Linear Elements for Improving the Step-Response of a Particular Servomechanism, R. Naylor. *M.S. Thesis*, Department of Electrical Engineering, Manchester University, Manchester, England, 1953.

5. The SERME Servo System, J. C. West. *British Patent 29011/51*, 1951.

6. The Frequency Response of a Certain Class of Non-Linear Feedback Systems, J. C. West, J. L. Douce. *British Journal of Applied Physics*, London, England, vol. 5, 1954, pp. 204–06.

7. The Behaviour of a Remote Position Control Servomechanism with Hardspring Non-Linear Characteristics, J. C. West, P. N. Nikiforuk. *Proceedings*, Institution of Electrical Engineers, London, England, vol. 101, pt. II, 1954, pp. 481–92.

8. Open-Loop Frequency Response Method for Nonlinear Servomechanisms, R. L. Cosgriff. *AIEE Transactions*, vol. 72, pt. II, Sept. 1953, pp. 222–25.

9. A Frequency Response Method for Analyzing and Synthesizing Contactor Servomechanisms, Ralph J. Kochenburger. *Ibid.*, vol. 69, pt. I, 1950, pp. 270–84.

10. The Response of Remote-Position-Control Systems with Non-Linear Characteristics to Step-Function and Random Inputs, J. C. West, P. N. Nikiforuk. *Proceedings*, Institution of Electrical Engineers, London, England, vol. 102, pt. II, 1955, pp. 575–95.

11. A Technique for Non-Linear Function Generation, P. N. Nikiforuk. *Electronic Engineering*, London, England, vol. 27, no. 325, 1955, pp. 118–19.

Comparison of Backlash and Hysteresis Effects in Second-Order Feedback Systems

L. M. VALLESE
ASSOCIATE MEMBER AIEE

Synopsis: The different behaviors of a second-order feedback control system of autonomous type (i.e., with no external drive), with backlash or with hysteresis in the feedback link of the loop, are investigated. Approximate expressions for the stability limits and for the amplitude and frequency of the oscillations are derived.

A PROCEDURE for the analysis of multivalued nonlinearities in autonomous second-order feedback systems is discussed in a recent paper.[1] It is shown that by application of the Kryloff-Bogoliuboff method[2] it is possible to obtain in a straightforward manner approximate expressions for the starting condition of oscillations, as well as for the Fourier series terms representing the oscillations. The accuracy of these results is not discussed here, as it is not directly pertinent to this paper. Instead, the approximate results obtained by application of this method when backlash or hysteresis occur in the system are used to obtain general insight as to the different effects produced by these two types of multivalued nonlinearities.

Analytical Representations of Backlash and Hysteresis Characteristics

Typical examples of hysteresis characteristics of a supermalloy and of a Deltamax sample are shown in Figs. 1 and 2 (as afforded by courtesy of the Arnold Engineering Company, Marengo, Ill.). These characteristics, though obtained with a 60-cycle sinusoidal magnetizing force $x=A_x \sin\omega t$, are assumed as practically independent of frequency in the limited low range of frequencies of interest in this paper. Comparison with a backlash characteristic (Fig. 3) shows that the fundamental difference between the two types of characteristics is in the dependence of the x- and y- intercepts upon A_x and in the presence of saturation effects. In other words, whereas for backlash the x- and the y-intercepts and the mean locus are, respectively,

$$x_0 = \text{constant},\ y_0 = kx_0 \tag{1}$$

$$y_1 = kx \tag{2}$$

for hysteresis the corresponding quantities are, respectively, for the supermalloy samples of Fig. 4

$$x_0 \cong \alpha A_x - \beta A_x^3,\ y_0 = kx_0 \tag{3}$$

$$y_1 \cong kx - Nx^3 \tag{4}$$

and for the Deltamax samples of Fig. 5

$$x_0 \cong \alpha A_x - (\alpha A_x - x_{0s})\, U(\alpha A_x - x_{0s}) \tag{5}$$

$$y_0 \cong kx_0 - k(x_0 - x_{0s})\, U(x_0 - x_{0s}) \tag{6}$$

$$|y_1| = k|x| - k(|x| - x_1)\, U(|x| - x_1) \tag{7}$$

In equations 5, 6, and 7, x_{0s} and x_1 are saturation values with $x_{0s}=\alpha x_1/(1-\alpha)$, and $U(\ )$ denotes the unit step function.

The loops of Fig. 1 can be considered, approximately, as "ellipses" developing around a mean curve given by equation 4. They may be represented, therefore, by the equation

$$y = y_1 + y_2 = kx - Nx^3 + k\left(\sqrt{1-\left(\frac{x_0}{A_x}\right)^2} - 1\right)x \pm kx_0\sqrt{1-\left(\frac{x}{A_x}\right)^2} \tag{8}$$

where it is assumed that $x=A_x\sin\psi$ and that the minus sign applies for $-\pi/2 \leq \psi \leq \pi/2$, the plus sign for $\pi/2 \leq \psi \leq 3\pi/2$. Equivalently, equation 8 may be written

$$y = kx - Nx^3 + k\left(\sqrt{1-\left(\frac{x_0}{A_x}\right)^2} - 1\right)A_x \sin\psi - kx_0\cos\psi = kA_x \sin(\psi - \psi_1) - N_x^3 \tag{9}$$

Similarly, the loops of Fig. 2 can be considered as parallelograms, developing around a mean curve given by equation 7. Their analytical representations are as follows: When no saturation occurs, i.e. for $A_x \leq x_1/(1-\alpha)$, the representation is

$$y = kx + f_1(x) \tag{10}$$

where

$f_1(x) =$	for
$-kx_0$	$\psi_1 \leq \psi \leq \psi_2$
$k(x-A_x)$	$\psi_2 \leq \psi \leq \pi/2$
0	$\pi/2 \leq \psi \leq \pi - \psi_2$
$k(A_x - x_0 - x)$	$\pi - \psi_2 \leq \psi \leq \pi + \psi_1$
kx_0	$\pi + \psi_1 \leq \psi \leq \pi + \psi_2$
$k(x+A_x)$	$\pi + \psi_2 \leq \psi \leq 3\pi/2$
0	$3\pi/2 \leq \psi \leq 2\pi - \psi_2$
$-k(A_x - x_0 + x)$	$2\pi - \psi_2 \leq \psi \leq 2\pi + \psi_1$

In these expressions

$$\sin\psi_1 = 2x_0/(A_x) - 1;\qquad \sin\psi_2 = 1 - x_0/A_x \tag{11}$$

When saturation occurs, i.e., $A_x \geq x_1/(1-\alpha)$, the representation is

$$|y| = k|x| - k(x - x_1)\, U(x - x_1) + |f_2(x)| \tag{12}$$

where

$f_2(x) =$	for
	$\psi_1 \leq \psi \leq \psi_2$
$k[x - x_1/(1-\alpha)]$	$\psi_2 \leq \psi \leq \psi_3$
0	$\psi_3 \leq \psi \leq \pi - \psi_3$
$k(x_1 - x)$	$\pi - \psi_3 \leq \psi \leq \pi + \psi_1$
$kx_1/(1-\alpha)$	$\pi + \psi_1 \leq \psi \leq \pi + \psi_2$
$k(x + x_1/(1-\alpha))$	$\pi + \psi_2 \leq \psi \leq \pi + \psi_3$
0	$\pi + \psi_3 \leq \psi \leq 2\pi - \psi_3$
$-k(x_1 + x)$	$2\pi - \psi_3 \leq \psi \leq 2\pi + \psi_1$

In these expressions

$$\sin\psi_1 = \frac{2\alpha - 1}{1-\alpha}(x_1/A_x)$$

$$\sin\psi_2 = x_1/A_x \tag{13}$$

$$\sin\psi_3 = \frac{1}{1-\alpha}(x_1/A_x)$$

Because of their different forms, equations 9, 10, and 12 must be treated separately. With reference to equation 9, for the purpose of comparison it is convenient to assume that the same equation represents a fictitious backlash characteristic, provided $N=0$ and $y_0 = \text{constant}$. Similarly equation 10 represents a backlash characteristic, if $x_0 = \text{constant}$ and A_x is not restricted (i.e., $x_1 = \infty$).

Solution

The differential equation which describes the behavior of a second-order feedback control system with no external drive and with the multivalued non-

Paper 56-454, recommended by the AIEE Feedback Control Systems Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE–ASME Nonlinear Controls Conference, Princeton, N. J., March 26–28, 1956. Manuscript submitted October 20, 1955; made available for printing February 14, 1956.

L. M. VALLESE is with Polytechnic Institute of Brooklyn, Brooklyn, N. Y.

This work was supported by the Office of Naval Research under contract with the Microwave Research Institute of the Polytechnic Institute of Brooklyn. The author gratefully acknowledges the encouragement and advice received from Dr. E. Weber.

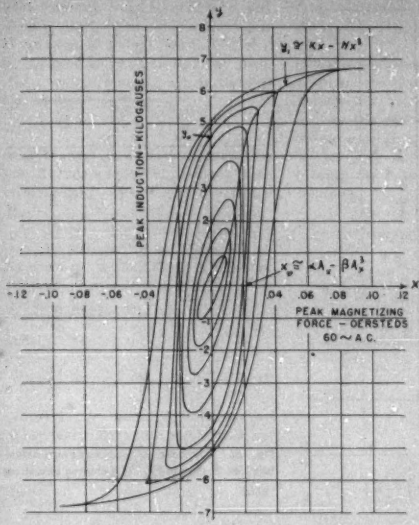


Fig. 1. Hysteresis characteristics of a supermalloy sample

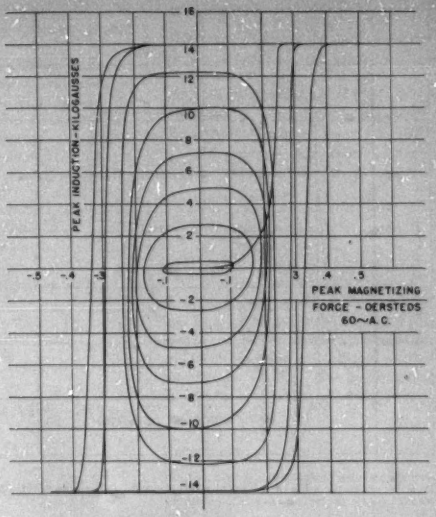


Fig. 2. Hysteresis characteristics of a Deltamax sample

linearity in the feedback link of the loop is

$$\ddot{x}+a_1\dot{x}+a_2y(x)=0 \tag{14}$$

For small values of a_1, assuming that the nonlinearity of equation 14 is small enough, an approximate solution of equation 14 can be obtained by means of the Kryloff-Bogoliuboff method. Thus, proceeding as detailed in reference 2, by taking

$$x=A_x\sin(\omega t+\varphi(t))=A_x\sin\psi,$$
$$\dot{x}=A_x\omega\cos\psi$$

where A_x and φ are assumed to vary "slowly" with time, provides the following approximate derivations of the amplitude and the frequency of oscillations

$$\frac{dA_x}{dt}=-\frac{1}{2\pi\omega}\int_0^{2\pi}(a_1\dot{x}+a_2y(x))\cos\psi d\psi \tag{15}$$

$$\frac{d\psi}{dt}=\Omega=\omega+\frac{1}{2\pi\omega A_x}\int_0^{2\pi}(a_1\dot{x}+a_2y(x))\sin\psi d\psi \tag{16}$$

The possible steady-state values of A_x corresponding to sustained oscillations are obtained by setting $dA_x/dt=0$; the stability of such oscillations is determined by note of the sign assumed by

$$\frac{d}{dA_x}\left(\frac{dA_x}{dt}\right)$$

for these values of A_x: the oscillations are stable or not, as the sign is negative or positive.

First, suppose that $y(x)$ in equations 15 and 16 is given by equation 9, and that it represents either a hysteresis characteristic (with x_0 given by equation 3) or a fictitious backlash characteristic (with x_0 given by equation 1, $N=0$). The amplitude and the frequency of oscillation are then given by the equations

$$\frac{dA_x}{dt}=\frac{-1}{2\omega}[a_1\Omega A_x-a_2kx_0] \tag{17}$$

and

$$\Omega=\frac{\omega}{2}\left[1+\sqrt{1-\left(\frac{x_0}{A_x}\right)^2}\right]-\frac{3Na_2A_x^2}{8\omega} \tag{18}$$

where $\omega=\sqrt{ka_2}$

In the case of backlash, sustained oscillations result for only one value of A_x. The corresponding characteristics are

$$A_x\cong\frac{ka_2}{a_1\Omega}x_0=\frac{\zeta a_1}{\Omega}x_0$$

$$\Omega\cong\frac{\omega}{2}\left[1+\sqrt{1-\frac{a_1\Omega}{ka_2}}\right]\cong\left(\sqrt{\zeta}-\frac{1}{4}\right)a_1\cong\omega-\frac{a_1}{4} \tag{19}$$

and

$$\frac{d}{dA_x}\left(\frac{dA_x}{dt}\right)=\frac{-1}{2\omega}\left[a_1\Omega+a_1A_x\frac{d\Omega}{dA_x}\right]$$

In these relations $\zeta=ka_2/a_1^2$ is a non-dimensional quantity. It is seen that

$$\frac{d}{dA_x}\left(\frac{dA_x}{dt}\right)<0$$

Since the value A_x given by equation 19 is always larger than zero, the quiescent position is stable and sustained oscillations occur when $A_x=x_0$. Correspondingly

$$\zeta=1\cong{}^5/_4$$

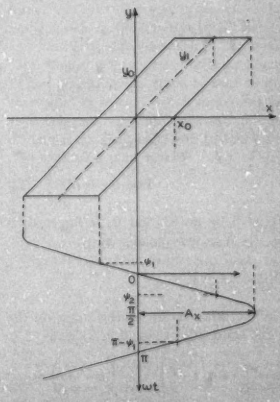


Fig. 3. Backlash characteristic for sinusoidal input

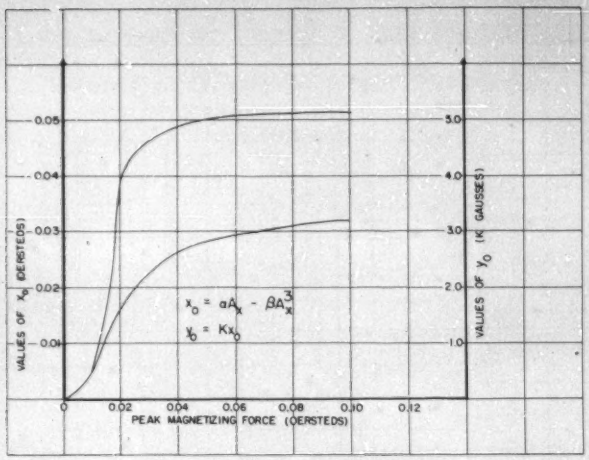


Fig. 4. Plot of x_0, y_0 versus A_x (according to Fig. 1)

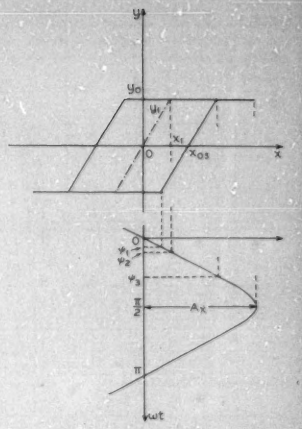


Fig. 5. Schematic of hysteresis characteristics (according to Fig. 2) including saturation

and in equation 9, $\psi_1 = \pi/_2$.

In summary, the system enters into a sustained stable oscillation if $\zeta \geq \sim 1/4$; both the amplitude and the frequency of this oscillation are functions of ζ.

In the case of hysteresis, sustained oscillations occur for two different solutions, one of which is $A_x = 0$

$$A_{x1} = 0; \quad A_{x2} = \sqrt{\frac{\alpha}{\beta} - \frac{a_1\Omega}{\beta k a_2}} \tag{20}$$

$$\Omega_1 = \frac{\omega}{2}\left[1 + \sqrt{1-\alpha^2}\right]$$

$$\Omega_2 = \frac{\omega}{2}\left[1 + \sqrt{1+\left(\frac{a_1\Omega}{k a_2}\right)^2}\right] - \frac{3Na_2}{8\omega}\left[\frac{\alpha}{\beta} - \frac{a_1\Omega}{\beta k a_2}\right] \tag{21}$$

and

$$\frac{d}{dA_x}\left(\frac{dA_x}{dt}\right) = \frac{-1}{2\omega}\left[a_1\Omega + a_1 A_x \frac{d\Omega}{dA_x} - k a_2 \alpha + 3\beta k a_2 A_x^2\right] \tag{22}$$

It is to be noted that Ω_2 is larger than ω. If N is very small, then

$$\Omega_2 \cong \omega\left(1 + \frac{1}{4\zeta}\right)$$

On the other hand, taking $\Omega_2 \cong \omega$ on the right-hand side of equation 21 gives

$$\Omega_2 = \frac{\omega}{2}\left[1 + \sqrt{1+\frac{1}{\zeta}}\right] - \frac{3Na_2}{8\omega\beta}\left(\alpha - \frac{1}{\sqrt{\zeta}}\right) = \omega\left(1 + \frac{1}{4\zeta}\right) - \frac{3Na_2}{8\omega\beta}\left(\alpha - \frac{1}{\sqrt{\zeta}}\right) \tag{23}$$

From equation 22 the rest position A_{x1} is stable or not, according as the following obtains

$$\zeta \lesssim \frac{\Omega_1}{\alpha a_1} \tag{24A}$$

On the other hand, neglecting $d\Omega/dA_x$ in equation 22 gives that the oscillation of amplitude A_{x2} is stable or not, according as

$$\zeta \gtrless \frac{\Omega_2}{\alpha a_1} \tag{24B}$$

Due to the approximations involved, if $\Omega_1 \cong \Omega_2 \cong \omega$ is taken in equations 24(A) and 24(B), the conclusion results that, according as

$$\sqrt{\zeta} \lesssim \sim \frac{1}{\alpha} \tag{25}$$

the stable position of the system is respectively that of rest or that of the oscillation of amplitude A_{x2}. Both A_{x2} and Ω_2 are functions of ζ; in particular the following obtains

$$A_{x2} = \sqrt{\frac{\alpha}{\beta} - \frac{\Omega_2}{\beta\zeta a_1}}$$

which reduces to $\sqrt{\alpha/\beta - 1/\beta\sqrt{\zeta}}$ if $\Omega_2 \cong \omega$.

Second, suppose that $y(x)$ in equations 15 and 16 is given by equation 10. This equation represents a hysteresis characteristic if $x_0 = \alpha A_x$, and x_1 is finite, or a backlash characteristic if $x_0 =$ constant and $x_1 = \infty$.

Solving equations 15 and 16 gives

$$\frac{dA_x}{dt} = \frac{-1}{2\pi\omega}\left[A_x(a_1\Omega\pi - 4x_0 k a_2) + 4x_0{}^2 k a_2\right] \tag{26}$$

$$\Omega = \omega\left[1 + \frac{1}{4\pi}(\sin 2\psi_1 - 6\sin 2\psi_2 + 2\psi_1 - 4\psi_2 + \pi)\right] \tag{27}$$

where ψ_1 and ψ_2 are given by equation 11.

In the case of backlash, the sustained oscillations are given by

$$A_x \cong x_0 \frac{2\zeta a_1}{\pi\Omega}\left[1 \pm \sqrt{1 - \frac{\pi\Omega}{\zeta a_1}}\right] \tag{28}$$

$$\Omega \cong \sqrt{\zeta a_1}\left[\frac{5}{4} + \frac{1}{4\pi}(2\psi_1 - 4\psi_2 + \sin 2\psi_1 - 6\sin 2\psi_2)\right] \tag{29}$$

and

$$\frac{d}{dA_x}\left(\frac{dA_x}{dt}\right) = \frac{-1}{2\pi\omega}\left(a_1\Omega\pi - 4x_0 k a_2 + \pi a_1 A_x \frac{d\Omega}{dA_x}\right) \tag{30}$$

Equation 28 is identical to equation 8(A) of reference 1, although the analytical representation of the backlash characteristic assumed here and in reference 1 are somewhat different.

In this case a critical value exists wherein the rest position is stable or not, if $\zeta \lesssim \zeta_{crit}$ and, correspondingly, a stable oscillation of finite amplitude occurs when $\zeta \geq \zeta_{crit}$. The ζ_{crit} value does not

follow directly from equations 28 and 29 on letting $A_x = 2x_0$. Graphical analysis of equation 30 shows[1] that $d/dA_x(dA_x/dt)$ changes to negative sign for $\zeta = 5$, $A_x = 1.6x_0$, but a more accurate step-by-step analysis[4] yields

$$\zeta_{crit} = 3.046$$

with $A_x = 2x_0$. It is to be noticed that both A_x and Ω vary with ζ, and that at the critical point $\zeta_1 = 0$, according to equation 11.

In the case of hysteresis, taking $x_0 = \alpha A_x$ in equations 28 and 29 and solving for sustained oscillations gives only one solution, of zero amplitude

$$A_{x1} \doteq 0, \ \Omega_1 \cong 2\omega$$

and

$$\frac{d}{dA_x}\left(\frac{dA_x}{dt}\right) \cong \frac{-1}{2\pi\omega}(a_1\Omega_1\pi - 4ka_2\alpha) \qquad (31)$$

provided $a_1\Omega_1\pi - 4ka_2\alpha \neq 0$. It follows that the system will be stable or not at the rest position according as

$$\zeta \lessgtr \frac{\pi\Omega_1}{4\alpha a_1} \qquad (32A)$$

thus, taking $\Omega_1 = 2\omega$, for the following equation, yields

$$\sqrt{\zeta} \lessgtr \frac{\pi}{2\alpha} \qquad (32B)$$

It is instructive to compare equations 32(A) and 32(B) with equations 24 and 25.

When the system is unstable at the rest position, any disturbance will result in passage into the saturation region. To analyze this case it is necessary to substitute equation 12 instead of equation 10 for $y(x)$ in equations 15 and 16. Solving the latter equations gives the following

$$\frac{dA_x}{dt} = \frac{-1}{2\pi\omega}\left(a_1\Omega_1 x_1\pi - \frac{4ka_2\alpha x_1^2}{(1-\alpha)A_x}\right)$$

$$\Omega = \omega\left\{1 + \frac{1}{2\pi}\left(\frac{2x_1}{A_x}\left[\frac{1-2\alpha}{1-\alpha}\cos\psi_1 - 2(1-\alpha)\cos\psi_1 + \frac{\cos\psi_2}{1-\alpha}\right]\right) + \frac{1}{2}(-\sin 2\psi_1 + 2\sin 2\psi_2 - \sin 2\psi_2 + \psi_1 - 2\psi_2 + \psi_1)\right\}$$

and

$$\frac{d}{dA_x}\left(\frac{dA_x}{dt}\right) = \frac{-1}{2\pi\omega}\left(2a_1\Omega\pi + a_1A_x\pi\frac{d\Omega}{dA_x}\right)$$

These relations yield that a stable oscillation is obtained when

$$A_x \cong 2x_1\sqrt{\frac{\zeta a_1}{\pi\Omega}}\sqrt{\frac{\alpha}{1-\alpha}}$$

Conclusions

The major differences in the behavior of the second-order system when backlash or hysteresis are present follow.

In the case of backlash, if $\zeta \geq \zeta_{crit}$, the system enters into a stable oscillation. The angular frequency Ω of this oscillation is generally less than $\omega = \sqrt{ka_2}$, as follows from equation 19

$$\Omega \leqq \omega - a_1/4$$

and from equation 28

$$\frac{\Omega}{\pi} \leqq \omega\frac{\sqrt{\zeta}}{\pi}$$

In the case of hysteresis, the rest position is stable or not according as $\zeta \lessgtr \zeta_{crit}$. In addition, there always occurs another oscillation which is unstable or not according as $\zeta \lessgtr \zeta_{crit}$. Its amplitude is limited by saturation; its frequency is, in general, larger than ω.

References

1. Analysis of Backlash in Feedback Control Systems With One Degree of Freedom, L. M. Vallese. *AIEE Transactions*, vol. 74, pt. II, Mar. 1955, pp. 1–4.

2. Introduction to Nonlinear Mechanics (book), N. Kryloff, N. Bogoliubeff. (English translation) Princeton University Press, Princeton, N. J., 1949.

3. Backlash in a Velocity Lag Servomechanism, Nathaniel B. Nichols. *AIEE Transactions*, vol. 72, pt. II, 1953 (Jan. 1954 section), pp. 462–67.

Describing Functions Representing the Effect of Inertia, Backlash, and Coulomb Friction on the Stability of an Automatic Control System—I

K. N. SATYENDRA
MEMBER AIEE

Synopsis: A first study is presented of the effect of backlash and coulomb friction upon the stability of an automatic feedback control system. The system consists of a mechanical linkage in which the load member is subjected to a restraining force. Describing functions that represent the load response for finite values of inertia, backlash, and coulomb friction between the load and an external surface are derived analytically. An example illustrates the conditions under which nonlinearities cause sustained oscillations in a given system, and how the Nyquist criterion can be used to obtain frequency and amplitude of the oscillations.

A WELL-KNOWN nonlinearity in feedback control systems is backlash between mechanical linkages used in the control system. Backlash effect on system stability has been investigated by Tustin[1] by the fundamental-component or the describing-function method. This paper presents an extension of Tustin's work and employs analytical rather than graphical methods of harmonic analysis. The equations of motion of the driving member and the driven member, in the presence of a finite value of backlash between them, are developed when the latter is subjected to different types of restraining forces such as coulomb friction or viscous damping. Under the assumption of a sinusoidal driving force, the load response is shown to be nonsinusoidal owing to the nonlinearity introduced by the backlash. The complex ratio of the fundamental component of this nonsinusoidal wave to that of the sinusoidal wave that would represent the load response were the backlash absent, is defined as the describing function for this particular nonlinearity.

Nomenclature

A_k, B_k = coefficients of cosine and sine terms of the kth harmonic; see equation 42

b = backlash between motor and load; see Fig. 1

Paper 56-453, recommended by the AIEE Feedback Control Systems Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE-ASME Nonlinear Controls Conference, Princeton, N. J., March 26–28, 1956. Manuscript submitted October 21, 1955; made available for printing February 27, 1956.

K. N. Satyendra is with the Westinghouse Electric Corporation, Baltimore, Md.

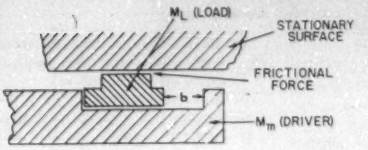


Fig. 1. Mechanical linkage consisting of driving motor and load with backlash and coulomb friction

Fig. 2. Characteristics of restraining force on load

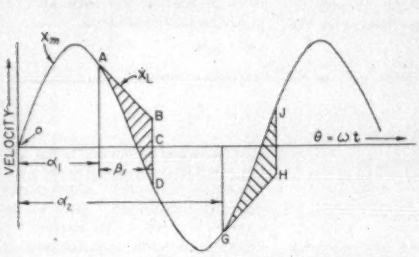


Fig. 3. Load response for sinusoidal input force (small values of λ)

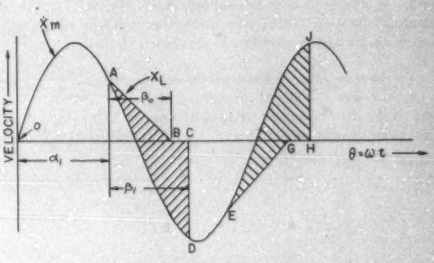


Fig. 4. Load response for sinusoidal input force (large values of λ)

b_L = effective load backlash; see equation 54
F = coulomb force per unit mass
M_L, M_m = masses of load and motor respectively
x, x_m, x_L = displacements of center of mass, motor, and load respectively
X, $\dot{X}$, X_{fL}, x_{fL}: see equations 1, 2, and 4
α_1, β_1, β_0, α_1', β_1': see Figs. 3, 4, and 5
φ_{fL}, ϕ_{fL} (2) = see equations 4, 6, and 7
$\lambda = b_L/X$ = ratio of load backlash to amplitude
$\nu = F/\omega^2 X$ = nondimensionalized force parameter
ω = angular frequency of sinusoidal motion
$C_1 = M_m/M_m + M_L$: see equation 54
C_3: see equation 13

System Under Study and Assumption

The translatory motion is considered of two finite masses M_m and M_L, representing the driving member and the driven member respectively of a mechanical linkage, as shown in Fig. 1. For ease of visualizing relative motion between the two masses, it will be assumed that the motor mass M_m is infinitely larger than the load mass M_L so that the presence of the driven member has no effect on the motion of the driving member. This assumption is somewhat analogous to that used in investigating the stability of two finite synchronous machines which are replaced, for purposes of analysis, by a finite synchronous machine swinging with respect to either an infinite bus or a synchronous machine of infinite inertia. The actual effect of finite motor mass can be readily introduced by identifying the motion of the hypothetical infinite mass as that of the center of mass of two masses M_m and M_L; see Appendix III.

It will be assumed that the driver of a very large mass M_m is made to execute simple harmonic motion by means of an externally applied force. The load mass M_L, separated from the motor mass by a backlash, b_L, is assumed not to affect the motion of the motor. Bouncing action between the motor and load when they strike each other is taken to be negligible.

Under these assumptions, the nature of the load motion will be governed by the type of restraining force acting on the load. The effect of a coulomb force F', having the characteristics shown in Fig. 2 can now be investigated. This is the force acting between the load member M_L and a stationary external surface. The effect of viscous damping and other restraints will be investigated subsequently.

Wave Form of Load Response Resulting From Sinusoidal Input

Let x_m and x_L respectively represent the instantaneous positions of the motor and load relative to a fixed point. Let the simple harmonic motion of the motor be defined by the equation

$$x_m = X \cos \omega t \tag{1}$$

in which X = amplitude, ω = angular frequency, and t = the time. The instantaneous velocity $\dot{x}_m$ of the motor will then be given by

$$\dot{x}_m = \dot{X} \sin \omega t \tag{2}$$

in which

$$\dot{X} = \omega X \tag{3}$$

In the absence of backlash, the load response would have been sinusoidal and the values of x_L and $\dot{x}_L$ at any instant of time could be represented by equations 1 and 2 respectively. Owing to the presence of backlash, however, the load response becomes nonsinusoidal. If the load response is periodic, it can be analyzed into its harmonics by the usual methods and the fundamental component x_{fL} of the load can be written as

$$x_{fL} = -X_{fL} \cos (\omega t + \phi_{fL}) \tag{4}$$

in which X_{fL} and ϕ_{fL} are the amplitude and phase of the fundamental component. The describing function representing the load response is defined symbolically by the equation

$$N(x_L) = \frac{X_{fL}}{X} / \phi_{fL} \tag{5}$$

Instead of considering displacements, it will be found more instructive to study the velocities of the motor and the load. The wave form representing load velocity can be analyzed into its harmonics so that its fundamental component $\dot{x}_{fL}$ can be written in the form

$$\dot{x}_{fL} = A_1 \cos \omega t + B_1 \sin \omega t \tag{6}$$

It has been shown in Appendix II, that

$$\begin{aligned} \dot{X}_{fL} &= \omega X_{fL} \\ \phi_{fL} &= \phi_{fL} \end{aligned} \tag{7}$$

Thus

$$N(x_L) = \frac{\dot{X}_{fL}}{\dot{X}} / \phi_{fL} \tag{8}$$

This definition of the describing function in terms of the amplitude and phase

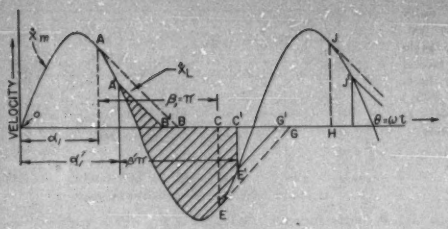


Fig. 5 (above). Load response for sinusoidal input force (extreme values of λ)

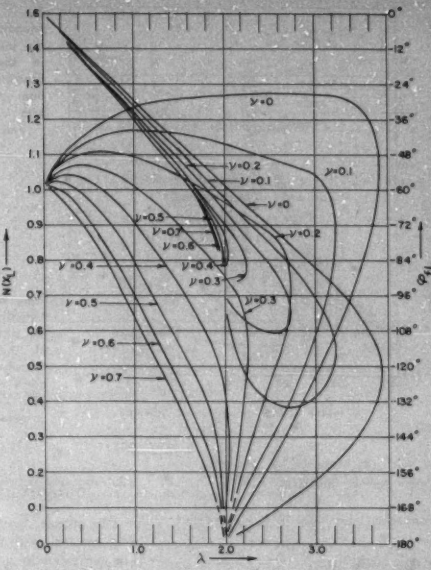


Fig. 6 (right). Magnitude and phase angle of the describing function $N(x_L)$ versus λ

of velocity rather than of displacement will be used since it simplifies the mathematics involved in the analysis. Therefore in the following analysis the load velocity rather than the load displacement is investigated for a number of cases which are identified in terms of a nondimensionalized parameter λ as defined by the equation

$$\lambda = \frac{b_L}{X} \tag{9}$$

CASE 1: SMALL VALUES OF λ

Let the sinusoidal curve of Fig. 3 represent the velocity of the motor. Let it be assumed that the motor and the load are in contact with each other at point 0. The load and the motor move with equal velocities over the portion $0A$ of this velocity curve. At A, the inertia force on the load just exceeds the restraining force and hence it is separated from the motor on the driving side. The two masses travel with different velocities until the backlash is closed on the nondriving side. The velocities of the two masses become equal at point D and they move together until the inertia force on the load exceeds the restraining force. This excess occurs at point G; the process then repeats itself.

CASE 2: LARGE VALUES OF λ

In Fig. 3, it is seen that the load velocity during separation decreases uniformly along AB. If the backlash b_L is large in comparison with X, the load velocity curve wil. appear as in Fig. 4. During the positive half-cycle, the velocity AB decreases uniformly until it becomes zero, as at point B. The load mass becomes stationary and remains in contact with the external stationary surface until the backlash is closed on the nondriving side. The two masses come into contact at D and move together with the same velocity along the curve DE. For certain values of λ, the process repeats itself periodically.

CASE 3: EXTREME VALUES OF λ

Referring to Fig. 4, it can be shown that λ gradually increases as the angle β_1 approaches π. If, when $\beta_1=\pi$, the amplitude X or backlash b_L is suddenly altered, the load velocity will undergo a transition from the dashed curve to the solid curve shown in Fig. 5. Physically, this transition means that the two masses separate (or come into contact) at A', a point which is so located that

$$\alpha_1' > \alpha_1 = \cos^{-1}(-\nu) \tag{10}$$

They come into contact (or separate) at E', a point which is separated from A' by π radians. The value of λ first increases, as α_1' increases, and then diminishes after reaching a maximum value.

Stability Considerations

GENERAL REMARKS

The describing functions derived in Appendix I can be plotted in different ways. In Fig. 6, the magnitude and phase angle of the describing function $N(x_L)$ are plotted against the ratio λ of load backlash to amplitude for a number of values of the nondimensionalized force parameter ν, as defined in Appendix I. Of particular interest is the fact that, for small values of $\nu = F/\omega^2 X$, the ratio $\lambda = b_L/X$ has a maximum value beyond which it decreases, reaching a final value of $\lambda = 2$. This means that for a fixed value of the load backlash b_L and a given coulomb force F, the amplitude decreases with increasing frequency up to a minimum value beyond which load response either ceases or becomes irregular. Conversely, for the same conditions, the amplitude increases with decreasing frequency, possibly giving rise to an unstable situation. In Fig. 7, the magnitude and phase angle of the inverse describing function $-1/N(x_L)$ in terms of λ and ν are plotted.

Situations exist in which the mechanical linkage in question may be part of a control system with feedback and may give rise to sustained oscillations of a definite frequency and amplitude. Whether or not such oscillations exist can be ascertained by the use of the Nyquist stability criterion, as subsequently outlined. Before discussing such applications of the describing function, it is desirable to plot these functions using frequency and amplitude as the variable parameters.

From equation 9, the amplitude X can be written, in terms of the effective load backlash b_L (see Appendix III), as

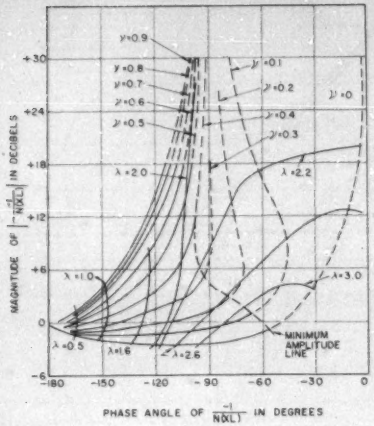


Fig. 7 (left). Magnitude versus phase angle of the inverse describing function $-1/N(X_L)$ with λ and ν as parameters

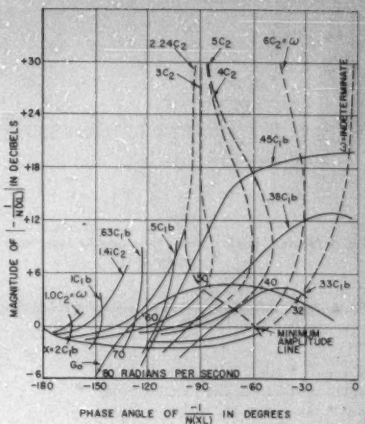


Fig. 8 (right). Magnitude versus phase angle of the inverse describing function $-1/N(X_L)$ with ω and X as parameters

$$X=\left(\frac{1}{\lambda}\right)b_L=\left(\frac{1}{\lambda}\right)bC_1 \qquad (11)$$

From equations 9 and 25, it follows that

$$\omega=\sqrt{\frac{F\lambda}{\nu b_L}}=\sqrt{\frac{F'\times M_m}{M_L(M_m+M_L)b}}\sqrt{\frac{\lambda}{\nu}}$$

$$=C_2\sqrt{\frac{\lambda}{\nu}} \qquad (12)$$

in which

$$C_2=\sqrt{\frac{F'}{M_L}\,\frac{M_m}{M_m+M_L}\,\frac{1}{b}} \qquad (13)$$

is a constant for a given mechanical system. By selecting those points in Fig. 7 for which the ratio $\lambda/2$ is constant, curves of constant frequency, Fig. 8, can be drawn.

Nyquist Plots

Let the example of a hypothetical control system be studied as made up of a linear element, represented by the transfer function $G_0(j\omega)$, in series with the nonlinearity, represented by the describing function $N(x_L)$, as shown in the block diagram of Fig 9. As pointed out by E. C. Johnson,[3] intersections between the phase-gain plots of $G_0(j\omega)$ and $-1/N(x_L)$ represent points at which the Nyquist stability criterion can be applied. If the frequencies of the two functions are equal at these points, convergent or divergent oscillations occur. The inverse describing function $-1/N(x_L)$ is shown in Fig. 8, with the frequency ω and amplitude X of the driving motor as the variable parameters. For the sake of illustration, the following transfer function is plotted on the same graph

$$G_0(s)=\frac{1}{(0.02s)^3+2(0.006)s+1} \qquad (14)$$

The inverse describing function which runs quite close to the plot of $G_0(s)$ is the one corresponding to $\omega=2.24C_2$. These two curves are closest for ω approximately equal to 60 radians per second. If $C_2>60/2.24$ or $C_2>27$ (i.e., large coulomb force or small backlash), there are no significant intersections between the two functions, and consequently the system is stable.

It is also seen that the curve $\omega=3C_2$ of $-1/N(x_L)$ runs very close to the point of intersection of G_0 and the minimum amplitude line, which occurs at approximately $\omega=50$ radians per second. Thus if C_2 lies between 27 and 16.7 (i.e., if the coulomb force is decreased or backlash increased), one or more significant intersections may occur between the plots of G_0 and $-1/N\ (x_L)$. These intersections will occur on the left-hand side of the "minimum amplitude line." Fig. 10 sketches two such intersections, A and B. It can be shown that point A is a low-frequency divergent-equilibrium point, while point B is a high-frequency convergent-equilibrium point. If the value of C_2 is decreased below 16.7 (i.e., if coulomb force is decreased or backlash increased further), significant intersections may occur on either side of the minimum amplitude line. Fig. 11 sketches two such intersections, C and D. It can be proved that C is a low-frequency convergent-equilibrium point, while D is a high-frequency convergent-equilibrium point. If coulomb force is reduced to zero (i.e., if $C_2=0$), there will be two convergent oscillations at approximate frequencies of 32 and 70 radians per second respectively.

These numerical examples illustrate how the engineer in automatic control systems like those considered in this paper can prevent the possibility of sustained oscillations by the introduction of the desired amount of coulomb force. The use of other types of restraining forces, such as viscous damping, will be the subject of a subsequent paper.

Conclusions

1. In a mechanical linkage consisting of a driving motor and a driven member, and forming part of a feedback control system, the presence of backlash between the two members of the mechanical linkage may lead to sustained oscillations.

2. If a restraining force, such as coulomb friction, is made to exist between the driven load and an external stationary surface, the possibility of sustained oscillations is diminished.

3. The stability of the automatic control system can be investigated by studying the Nyquist plot of the inverse describing function of the load response.

4. Sustained oscillations may be eliminated by increasing the value of the parameter C_2 defined in equation 13.

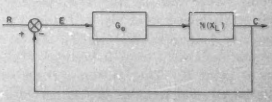


Fig. 9. Block diagram of linear element and nonlinear element representing backlash and coulomb friction

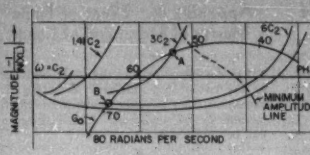


Fig. 10. Intersection of $G_0(j\omega)$ and $-1/N(x_L)$ showing convergent and divergent equilibrium points

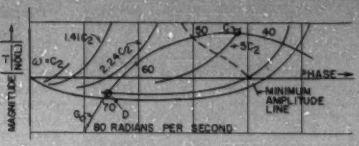


Fig. 11. Intersection of $G_0(j\omega)$ and $-1/N(x_L)$ showing two convergent equilibrium points

Appendix I. Derivation of Describing Functions

The equations of the load mass response will be derived for the three cases shown in Figs. 3, 4, and 5.

Case 1

Fig. 3 shows that, along $0A$

$$\dot{x}_L = \dot{x}_m = \omega X \sin \omega t, \text{ for } 0 \leq \theta \leq \alpha_1 \quad (15)$$

Along AB the velocity of the load is governed by the equation

$$M_L \ddot{x}_L + F' = 0 \quad (16)$$

Solution of equation 16 yields

$$\dot{x}_L = -\frac{F'}{M_L} t + c \quad (17)$$

in which c is a constant of integration. From equations 15, 16, and 17 it follows that where $\theta = \alpha_1$

$$\cos \alpha_1 = -\frac{F}{\omega^2 X} \quad (18)$$

$$c = \frac{F}{\omega} \alpha_1 + \omega X \sin \alpha_1 \quad (19)$$

in which F, the coulomb force per unit load mass, is given by

$$F = F'/M_L \quad (20)$$

During one half-cycle, the equation of the load velocity curve can be written as

$$\dot{x}_L = -\frac{F}{\omega}\theta + \omega X \sin \alpha_1 + \frac{F}{\omega}\alpha_1 \text{ for } \alpha_1 \leq \theta \leq (\alpha_1 + \beta_1)$$

$$= \omega X \sin \theta \text{ for } (\alpha_1 + \beta_1) \leq \theta \leq (\alpha_1 + \pi) \quad (21)$$

The curve is symmetric during the second half-cycle and satisfies the relation

$$\dot{x}_L \text{ at } \theta = -\dot{x}_L \text{at } (\theta + \pi) \quad (22)$$

The total distance travelled by the load relative to the motor, between points A and D is b_L. This distance is equal to the shaded area shown in Fig. 3. From equations 2 and 21

$$b_L = \frac{1}{\omega}\int_{\alpha_1}^{\alpha_1+\beta_1} (\dot{x}_L - \dot{x}_m) d\theta$$

$$= \frac{1}{\omega}\left[\int_{\alpha_1}^{\alpha_1+\beta_1} \left(-\frac{F}{\omega}\theta + \omega X \sin \alpha_1 + \frac{F}{\omega}\alpha_1\right) d\theta - \int_{\alpha_1}^{\alpha_1+\beta_1} \omega X \sin \theta d\theta\right]$$

$$= -\frac{1}{2}\frac{F}{\omega^2}\beta_1^2 + X[\beta_1 \sin \alpha_1 - \cos \alpha_1 \times \cos(\alpha_1 + \beta_1)] \quad (23)$$

From equation 23, it follows that

$$\lambda = \frac{b_L}{X} = -\frac{1}{2}\nu\beta_1^2 + \beta_1 \sin \alpha_1 - \cos \alpha_1 + \cos(\alpha_1 + \beta_1) \quad (24)$$

in which ν is a nondimensionalized parameter given by

$$\nu = \frac{F}{\omega^2 X} \quad (25)$$

For given values of ν and λ, the unknown angles α_1 and β_1 can be evaluated by solving equations 24 and 18.

Case 2

The equation of the straight line AB of Fig. 4 can be written as

$$\dot{x}_L = -\frac{F}{\omega}\theta + \omega X \sin \alpha_1 + \frac{F}{\omega}\alpha_1 \quad (26)$$

At point B, corresponding to

$$\theta = \alpha_1 + \beta_0 \quad (27)$$

it is seen that the velocity is zero. Hence from equations 26 and 27

$$\beta_0 = \frac{\omega^2 X}{F} \sin \alpha_1 = \frac{1}{\nu} \sin \alpha_1 \quad (28)$$

The velocity curve $\dot{x}_L$ can be written as

$$\dot{x}_L = -\frac{F}{\omega}\theta + \omega X \sin \alpha_1 + \frac{F}{\omega}\alpha_1, \text{ for } \alpha_1 \leq \theta \leq (\alpha_1 + \beta_0)$$

$$= 0 \text{ for } (\alpha_1 + \beta_0) \leq \theta \leq (\alpha_1 + \beta_1)$$

$$= \omega X \sin \theta, \text{ for } (\alpha_1 + \beta_1) \leq \theta \leq (\alpha_1 + \pi) \quad (29)$$

The angle β_1 can be evaluated by equating the shaded area of Fig. 4 to b_L. Hence

$$b_L = \frac{1}{\omega}\left[\int_{\alpha_1}^{\alpha_1+\beta_1} \left(-\frac{F}{\omega}\theta + \omega X \sin \alpha_1 + \frac{F}{\omega}\alpha_1\right) d\theta - \int_{\alpha_1}^{\alpha_1+\beta_1} \omega X \sin \theta d\theta\right]$$

$$= -\frac{1}{2}\frac{F}{\omega^2}\beta_1^2 + X[\beta_1 \sin \alpha_1 - \cos \alpha_1 + \cos(\alpha_1 + \beta_1)] \quad (30)$$

For given values of ν, the angle β_0 can be evaluated from equations 18 and 28. For given values of ν and λ, the angle β_1 can be readily evaluated from equation 30.

Case 3

By analogy with equation 21, the equation of the curve $A'B'C'E'$ of Fig. 5 can be written as

$$\dot{x}_L = -\frac{F}{\omega}\theta + \omega X \alpha_1', \text{ for } \alpha_1' \leq \theta \leq (\alpha_1' + \beta_0')$$

$$= 0 \text{ for } (\alpha_1' + \beta_0') \leq \theta \leq (\alpha_1' + \pi) \quad (31)$$

in which

$$\beta_0' = \frac{1}{\nu} \sin \alpha_1' \quad (32)$$

and the angle α_1' satisfied equation 10. For this condition, the value of λ can be written as

$$\lambda = \frac{b_L}{X} = -\frac{1}{2}\nu(\beta_0')^2 + \beta_0' \sin \alpha_1' - \cos \alpha_1' + \cos(\alpha_1' + \pi)$$

A simple way of solving this equation is to evaluate λ for values of $\alpha_1 \leq \alpha_1' \leq \pi$ in 5-degree steps, and tabulate the values of λ versus α_1'. For any given value of λ, the angle α_1' can then be readily estimated by interpolation.

Fourier Analysis Leading to Describing Function

The wave forms of $\dot{x}_L$, shown in Figs. 3, 4, and 5 are periodic and exhibit full-wave symmetry as defined by equation 22. The coefficients A_1 and B_1 of equation 6 can be written as

$$A_1 = \frac{2}{\pi\omega}\int_{\theta}^{\theta+\pi} \dot{x}_L \cos \theta d\theta$$

$$B_1 = \frac{2}{\pi\omega}\int_{\theta}^{\theta+\pi} \dot{x}_L \sin \theta d\theta \quad (33)$$

CASE 1: SMALL λ OR $\beta_1 \leq \beta_0 < \pi$

From equations 21 and 33 the following equation is obtained

$$A_1 = \frac{2}{\omega\pi}\left[\int_{\alpha_1}^{\alpha_1+\beta_1} \left(-\frac{F}{\omega}\theta + \omega X \sin \alpha_1 + \frac{F}{\omega}\alpha_1\right) \times \cos \theta d\theta + \int_{\alpha_1+\beta_1}^{\alpha_1+\pi} \omega X \sin \theta \cos \theta d\theta\right]$$

$$= \frac{2X}{\pi}\left\{-\frac{1}{2} + \frac{1}{4}\cos 2\alpha_1 - \frac{1}{4}\cos(2\alpha_1 + 2\beta_1) + \sin \alpha_1 \sin(\alpha_1 + \beta_1) - \nu[\beta_1 \sin(\alpha_1 + \beta_1)\cos(\alpha_1 + \beta_1) - \cos \alpha_1]\right\} \quad (34)$$

$$B_1=\frac{2}{\omega\pi}\left[\int_{\omega_1}^{\alpha_1+\beta_1}\left(-\frac{F}{\omega}\theta+\omega X\sin\alpha_1+\frac{F}{\omega}\alpha_1\right)\times\sin\theta\,d\theta+\int_{\alpha_1+\beta_1}^{\alpha_1+\pi}\omega X\sin^2\theta\,d\theta\right]$$

$$=\frac{2X}{\pi}\left[\frac{1}{2}(\pi-\beta_1)+\frac{1}{4}\sin 2\alpha_1-\sin\alpha_1\cos(\alpha_1+\beta_1)+\frac{1}{4}\sin(2\alpha_1+2\beta_1)+\nu[\beta_1\cos(\alpha_1+\beta_1)-\sin(\alpha_1+\beta_1)+\sin\alpha_1]\right] \quad (35)$$

CASE 2: LARGE λ, OR $\beta_0\leq\beta_1<\pi$

From equations 29 and 33

$$A_1=\frac{2X}{\pi}\left\{-\frac{1}{2}+{}^1/_4\cos 2\alpha_1+\frac{1}{4}\cos(2\alpha_1+2\beta_1)+\sin\alpha_1\times\sin(\alpha_1+\beta_0)-\nu[\beta_0\sin(\alpha_1+\beta_0)+\cos(\alpha_1+\beta_0)-\cos\alpha_1]\right\} \quad (36)$$

$$B_1=2X/\pi\left\{\frac{1}{2}(\pi-\beta_1)+\frac{1}{4}\sin 2\alpha_1-\sin\alpha_1\cos(\alpha_1+\beta_0)+\frac{1}{4}\sin(2\alpha_1+2\beta_1)+\nu[\beta_0\cos(\alpha_1+\beta_0)-\sin(\alpha_1+\beta_0)+\sin\alpha_1]\right\} \quad (37)$$

CASE 3: EXTREME VALUES OF λ, OR $\beta_0<\beta_1=\pi$, $\alpha_1'>\alpha_1$

From equations 31 and 33

$$A_1=\frac{2X}{\pi}\left\{-\frac{1}{2}+\frac{1}{2}\cos 2\alpha_1'+\sin\alpha_1'\sin(\alpha_1'+\beta_0')-\nu[\beta_0'\sin(\alpha_1'+\beta_0')+\cos(\alpha_1'+\beta_0')-\cos\alpha_1']\right\} \quad (38)$$

$$B_1=\frac{2X}{\pi}\left\{-\frac{1}{2}\sin 2\alpha_1'-\sin\alpha_1'\cos(\alpha_1'+\beta_0')+\nu[\beta_0'\cos(\alpha_1'+\beta_0')-\sin(\alpha_1'+\beta_0')+\sin\alpha_1']\right\} \quad (39)$$

It must be pointed out that the equations presented in this paper are valid for all values of ν, and hence for $\nu=0$. This special case representing backlash between a motor and a load with no restraining force on the load corresponds to the first case treated by Tustin[1] and referred to by him as the case of zero friction.

Appendix II. Equivalence of Describing Functions of Displacement and Velocity

Consider the mechanical linkage of Fig. 1. In the absence of backlash, the load always moves along with the motor. Hence the displacement x_L and the velocity $\dot{x}_L$ of the load will be the same as those of the motor and can be given by the equations

$$x_L=-X\cos\omega t \quad (40)$$

$$\dot{x}_L=\omega X\sin\omega t \quad (41)$$

When backlash is present between the motor and the load, the load response becomes nonsinusoidal. Let the load velocity $\dot{x}_L$ be given by the Fourier series

$$\dot{x}_L=\sum_{k=1}^{\infty}\omega(A_k\cos k\omega t+B_k\sin k\omega t) \quad (42)$$

The steady-state displacement x_L may be obtained by integrating equation 42, term by term, to yield

$$x_L=\frac{A_k}{k}\sin k\omega t-\frac{B_k}{k}\cos k\omega t+C_k \quad (43)$$

in which C_k is a constant of integration. The fundamental components $\dot{x}_{fL}$ and x_{fL} of the velocity and the displacement are given respectively by the two following equations

$$\dot{x}_{fL}=\omega A_1\cos\omega t+\omega B_1\sin\omega t=\omega\sqrt{A_1^2+B_1^2}\sin\left(\omega t+\tan^{-1}\frac{A_1}{B_1}\right) \quad (44)$$

$$x_{fL}=A_1\sin\omega t-B_1\cos\omega t=-\sqrt{A_1^2+B_1^2}\cos\left(\omega t+\tan^{-1}\frac{A_1}{B_1}\right) \quad (45)$$

From equations 40 and 45, the describing function representing load displacement is written symbolically as

$$N(x_L)=\frac{\sqrt{A_1^2+B_1^2}}{X}\underline{/\tan^{-1}\frac{A_1}{B_1}} \quad (46)$$

From equations 41 and 44, the describing function representing the velocity is written symbolically as

$$N(\dot{x}_L)=\frac{\sqrt{A_1^2+B_1^2}}{X}\underline{/\tan^{-1}\frac{A_1}{B_1}} \quad (47)$$

It follows from equations 46 and 47 that

$$N(x_L)=N(\dot{x}_L) \quad (48)$$

Appendix III. Effect of Finite Masses of Motor and of Load

Let x = the distance of the center of mass of the two masses M_m and M of Fig. 1, measured from the fixed point from which x_m and x_L are measured. This distance x can be obtained as

$$x=\frac{M_m x_m+M_L x_L}{M_m+M_L} \quad (49)$$

From equation 49, the momentum equation becomes

$$M_m(\dot{x}_m-\dot{x})+M_L(\dot{x}_L-\dot{x})=0 \quad (50)$$

The relative displacement b of the two masses during separation for β_1/ω seconds can be written as

$$b=\int_0^{\beta_1/\omega}(\dot{x}_L-\dot{x}_m)dt \quad (51)$$

During the same interval of time, the displacements b_m and b_L of the motor and load respectively are

$$b_m=\int_0^{\beta_1/\omega}(\dot{x}-\dot{x}_m)dt \quad (52)$$

$$b_L=\int_0^{\beta_1/\omega}(\dot{x}_L-\dot{x})dt \quad (53)$$

From equations 50, 51, 52, and 53, it follows that

$$b=\int_0^{\beta_1/\omega}[(\dot{x}_L-\dot{x})+(\dot{x}-\dot{x}_m)]dt=\int_0^{\beta_1/\omega}[(\dot{x}_L-\dot{x})+M_L/M_m(\dot{x}_L-\dot{x})]dt=\left(\frac{M_L+M_m}{M_m}\right)b_L=\left(\frac{1}{c_1}\right)b_L \quad (54)$$

If $M_m>>M_L$, it follows from equation 54 that

$$b\simeq b_L \quad (55)$$

When the masses M_m and M_L are finite, the backlash b can be replaced by the load backlash b_L which is defined as the displacement of the load relative to the center of mass.

References

1. THE EFFECTS OF BACKLASH AND OF SPEED DEPENDENT FRICTION ON THE STABILITY OF CLOSED-CYCLE CONTROL SYSTEMS, A. Tustin. *Journal*, Institution of Electrical Engineers, London, England, vol. 94, pt. II, 1947, pp. 143–51.

2. SINUSOIDAL ANALYSIS OF FEEDBACK-CONTROL SYSTEMS CONTAINING NONLINEAR ELEMENTS, E. Calvin Johnson. *AIEE Transactions*, vol. 71, pt. II, July 1952, pp. 169–81.

Discussion

Abraham Fuchs (Westinghouse Electric Corporation, Baltimore, Md.): Dr. Satyendra has presented an excellent work extending our knowledge in the field of nonlinear servomechanisms on a problem which forms an essential element of many servomechanism designs.

In 1952, G. L. Turin in an unpublished thesis included derivations of some of the describing functions of Dr. Satyendra's paper. He also provided a time solution for certain cases which revealed that some intersections with the backlash describing function were false and did not represent system oscillations.

Tustin, in reference 1 of the paper, states that the worst case for stabilizing a system with backlash occurs when the output velocity does not change direction; i.e., its average velocity is high enough that its oscillations never cause the velocity to reverse. This will be called the constant velocity case. Dr. Satyendra's describing functions apply when the average output velocity is zero. This will be called the zero velocity case. Prof. Tustin's describing functions apply to the constant velocity case.

Comparing Dr. Satyendra's describing function, Fig. 7, with Prof. Tustin's Fig. 10, one finds that a given value of λ has more phase shift at constant speed than at zero speed. Thus the system at constant speed

must have a lower value of backlash to be stable than one at zero speed. To be certain, then, that a system with backlash is stable, it should be checked at a high operating speed as well as at zero speed. Note too that it is the running friction level which is critical and not the stiction level.

When the load friction is zero, almost any realizable system transfer function will produce an oscillation. Thus stabilization of such systems must depend upon a second nonlinearity such as dead space in the feedback element, error sensor, error amplifier, or in the motor response.

One very important conclusion which can be drawn from Dr. Satyendra's work and that of Prof. Tustin is that the approximation of backlash by the hysteresis describing function or by a hysteresis function generator in an analogue computer is totally inadequate. (For example, Greif[1] lumps hysteresis, backlash, and free-play together.) Hysteresis having the maximum of a 90-degree phase shift in its describing function can be stabilized when the low friction backlash system which it is representing cannot be so stabilized. The key role played by the output friction as a stabilizing influence is also missed by the hysteresis representation.

It is hoped that the author will follow this paper with transient response demonstrations of the significance of the intersections, comparing the actual system performance with that predicted by his describing functions.

REFERENCE

1. DESCRIBING FUNCTION METHOD OF SERVOMECHANISM ANALYSIS APPLIED TO MOST COMMONLY ENCOUNTERED NONLINEARITIES. H. D. Greif. *AIEE Transactions*, vol. 72, pt. II, Sept. 1953, pp. 243–48.

Distributing Large Single-Phase Loads in Limited-Capacity 3-Phase Power Systems

GEORGE F. AROYAN
STUDENT MEMBER AIEE

KEITH G. LAKEY
NONMEMBER AIEE

A LARGE single-phase load drawn from a limited-capacity 3-phase system will unbalance the 3-phase system. If analyzed by symmetrical components in the 3-phase system, equal and opposite positive- and negative-sequence currents are realized. This unbalance leads to excessive heating in the stator of the alternator, thus substantially reducing its over-all capacity. If uncorrected, the unbalance may also cause local hot spots in ordinary induction motors.

The methods of balancing the 3-phase load employ various types of large rotating machinery. Static methods using various transformer arrangements, unsymmetrical taps, etc., have been used, but none of these methods eliminate the negative-sequence current which is the cause of the unbalance.

A tapped reactor and capacitor have been used on electric welding machines in Great Britain;[1] and to balance a large single-phase load to 3-phase aircraft inverters.[2] In these two applications, however, the units were designed to balance a particular single-phase load condition, and any departure from this loading resulted in unbalance of the system. This paper presents a method developed to alter automatically these reactances irrespective of the magnitude, pf (power factor), or change in the single-phase load, and thus to convert continuously the single-phase load into a balanced 3-phase load. In addition, the pf of the load is improved to 0.866 when the method is applied.

Paper 56-625, recommended by the AIEE Industrial Power Systems Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Summer and Pacific General Meeting, San Francisco, Calif., June 25–29, 1956. Manuscript submitted February 29, 1956; made available for printing April 12, 1956.

GEORGE F. AROYAN is at the University of California, Berkeley, Calif., and LT. KEITH G. LAKEY, U. S. Navy, is at the San Francisco Naval Shipyard, San Francisco, Calif.

The work described was performed by the authors in partial fulfillment of the requirements for the degree of Naval Engineer at the Massachusetts Institute of Technology, and was supported by the U. S. Navy, Bureau of Ships. The authors wish to thank Dr. D. P. Campbell, professor of electrical engineering, Massachusetts Institute of Technology, for his encouragement and guidance as thesis adviser.

General Analysis

Fig. 1 is a schematic diagram of the basic power elements of the converter unit. The power elements consist of a capacitor and a tapped reactor connected across the 3-phase system. The three phases are *A*, *B*, *C*, and the single-phase load is connected across line *BC*. The midpoint connection between capacitor and inductance in *P*. The currents I_A, I_B, and I_C are the reflected 3-phase currents into the power source. The load current is I_L. The current through the capacitor is I_C. I_{C1} and I_{C2} are the division of this capacitive current through the inductive reactance. The current I_M is the current through the reactor induced by the line voltage V_{AB}. If X_C is the magnitude of capacitive reactance, X_{L1} the value of the inductive reactance between *P* and *B*, and X_{L2} the inductive reactance between *P* and *B*, the following relations hold

$$I_A = -I_M - I_{C1} \tag{1}$$

$$I_B = I_M - I_L - I_{C2} \tag{2}$$

$$I_C = I_L + I_{C1} + I_{C2} \tag{3}$$

and

$$jX_{L1} = \frac{V_{AP}}{I_M + I_{C1}} \tag{4}$$

$$jX_{L2} = \frac{V_{PB}}{I_M - I_{C2}} \tag{5}$$

$$-jX_C = \frac{V_{PC}}{I_{C1} + I_{C2}} \tag{6}$$

Fig. 2 is a vector diagram of the graphical analysis carried out for various values of load pf as the variable. The figure indicates the change of currents and voltages required for the single-phase load to produce balanced 3-phase currents. The reactances were normalized with respect to the load current and line voltage. For any particular pf, the reactances, phase currents, etc., vary directly with the magnitude of the load current. Under balanced conditions, the 3-phase currents are equal and in phase with their respective line-to-line voltages.

The following condition holds, relating the balanced 3-phase power to the single-phase load, assuming perfect conversion,

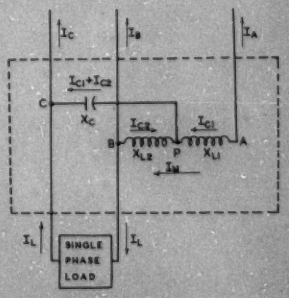


Fig. 1. Schematic diagram of basic static converter

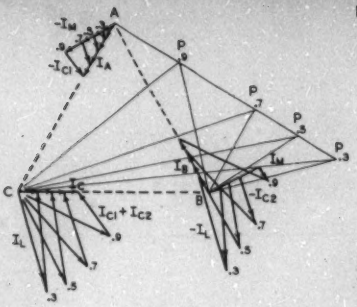
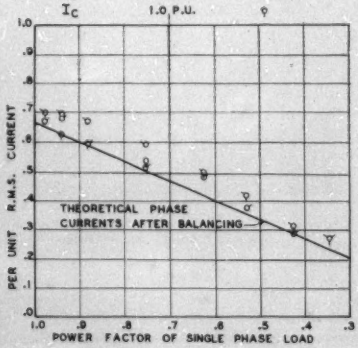


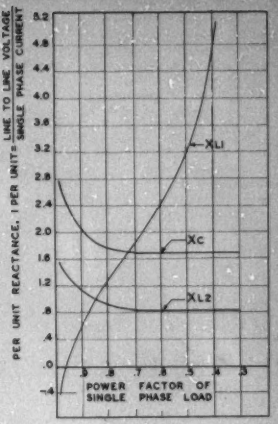
Fig. 2 (left). Vector diagram of graphical analysis

Fig. 3 (below left). Plot of observed phase currents for manually operated reactances

PHASE CURRENTS	BEFORE BALANCING	AFTER BALANCING
I_A	0.0 P.U.	α
I_B	1.0 P.U.	ο
I_C	1.0 P.U.	ο

Fig. 4 (right). Required reactances to convert single-phase load to balanced three phases





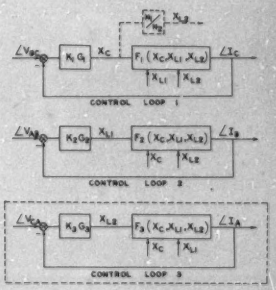


Fig. 5 (right). Control system block diagram

no reactance losses, leakage, distortion, etc.

$$3E_\phi I_\phi \cos 30° = E_L I_L \cos\theta \quad (7)$$

$$I_\phi = \frac{2}{3} I_L \cos\theta \quad (8)$$

where

E_ϕ = phase voltage
I_ϕ = balanced 3ϕ current
I_L = single-phase load
E_L = line voltage
$\cos\theta$ = Pf of single-phase load

The latter relationship is plotted in Fig. 3 as the "theoretical" value of the phase currents.

Fig. 3 is a plot of the phase currents obtained by using particular reactances for particular loads and shows correspondence between actual observations and the graphical analysis. Fig. 4 is a plot of the required per-unit reactance to balance the single-phase load obtained from the graphical analysis.

Control System

A feedback control system was used to permit self-adjustment of the reactances. The change in reactances as noted in Fig. 4 is linear with respect to load current magnitude, but nonlinear with respect to the pf of the load. From a study of the vector diagram, Fig. 2, a convenient measure of balanced conditions was observed to be the point at which the 3-phase currents are in phase with the line voltages. This phase measurement was used as the error measurement of the control system. One advantage of a feedback control is that an exact or precise operation on an input is not required. If stability conditions are satisfied, it is sufficient that the control system merely act in such a manner as to reduce the error. This basically is behind the operation of the feedback loops in this particular application. It was noted from an inspection of Fig. 2, that a change in the magnitude of I_C (i.e., X_C to a first approximation) had a pronounced effect on the phase-angle of I_C. Similarly, phase errors in I_A and I_B could be used to change the reactances X_H and X_L in such a way as to reduce the respective phase errors. At any one instant, a change in any one reactance gives rise to a change in all three phases, but as long as all three reactances are acting in the direction to reduce the error, a solution of balanced conditions eventually results with balanced magnitudes and zero phase error. The basic control loops for these three conditions are shown in Fig. 5 for separately controlled reactances. If a tap-changing transformer had been used in lieu of the separate inductances, the phase error of I_A could have been used to position the tap, and the phase error of I_B to change the total inductance between A and B.

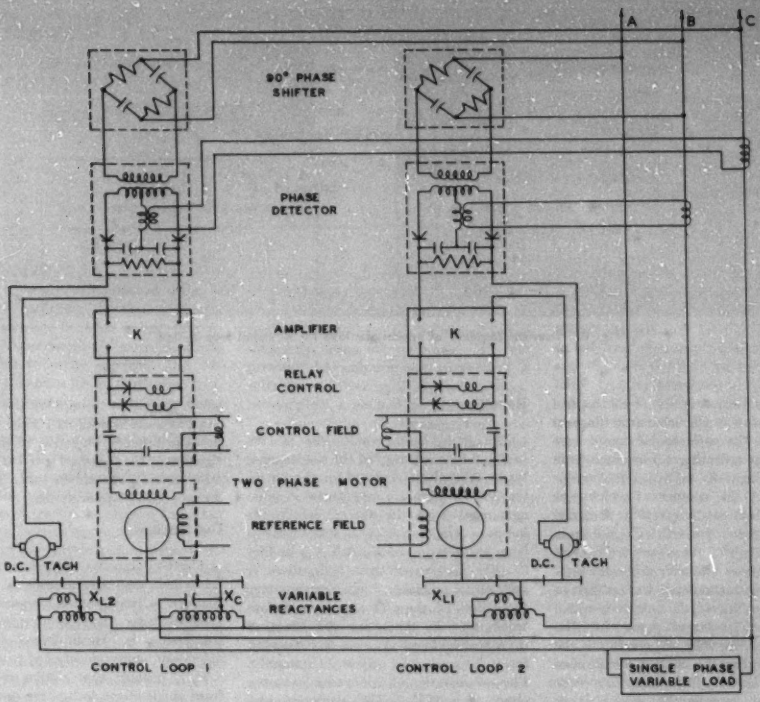


Fig. 6. Schematic wiring diagram for 2-control loop system

The system was initially investigated for correction of low-pf single-phase loads. In this particular application it may be noted from the reactance curves of Fig. 4 that the ratio of X_C and X_{L1} are proportional for pf less than about 0.75. Thus, to permit economy of units, the third control loop of Fig. 5 was eliminated. The reactance X_{L2} was determined from a direct ratio of the value of X_C. The elimination of this third control loop had no appreciable effect on the results for pf less than 0.75, but introduced some error into higher pf loads. All three control loops are recommended if a universal or random pf may be expected in the single-phase load. If the single-phase load substantially operates at only one pf, only one control loop is required and, of course, if there is no variation in magnitude or pf of the single-phase load, then invariant reactances and no control loops may be used to achieve balanced 3-phase distribution of the load.

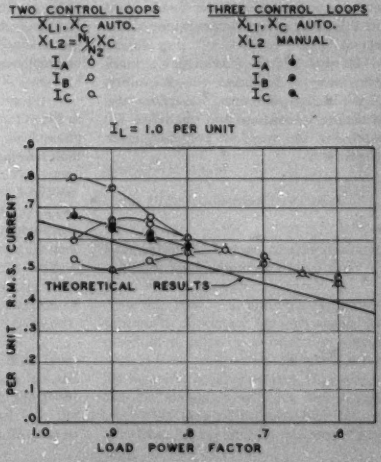


Fig. 7. Steady-state results for 2- and 3-control loop system

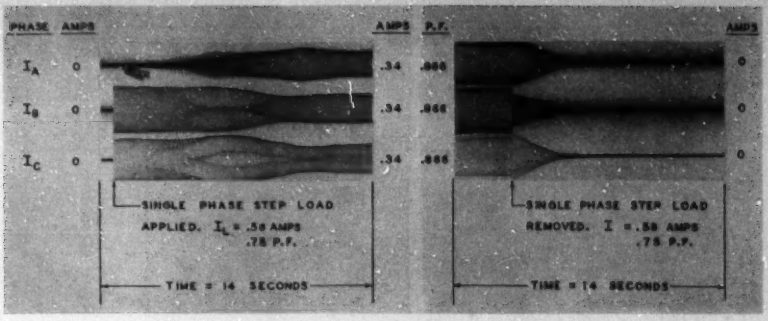


Fig. 8. Transient response of typical stepload for 2-control loop system

The actual construction of the control units is shown in the schematic diagram of Fig. 6. The variable reactances were obtained by reflecting a reactance from the secondary of an adjustable autotransformer, the contact of which was positioned by a small 2-phase servomotor. The servomotor received its signal from the phase error of the respective control loop through an amplifier and relay control. A d-c tachometer was added to provide stability of the relay-controlled servomotor. In larger power installations, it is possible to eliminate the amplifier, but at the small power level of the investigation, the amplifier was required to actuate the relay. It is possible to improve the servomotor response by providing dynamic braking with the relay control, and to eliminate servomotor excitation under steady-state conditions by proper arrangement of the relays.

The object of the investigation, however, was to determine the feasibility of the design and not to consider the extensive techniques and methods for improving the speed of response, error criteria, etc.

Results

The steady-state results are plotted in Fig. 7 for various pf of the single-phase load. The phase currents for the 2-control loop system are seen to be in good agreement with the theoretical results for pf of less than 0.75, as was surmised from the relation of X_C and X_{L1} in Fig. 3. For pf greater than 0.75, there is still an appreciable improvement over the values of the phase currents before balancing. To determine the effect of three control loops, X_{L1} was disconnected from X_C and was adjusted manually. The sole criterion of this manual was the phase error of I_A. The correction was thus applied in the same manner as an actual third control loop. The other two control loops were permitted to adjust their respective reactances X_C and X_{L1} automatically. These results are also plotted in Fig. 7, and indicate good correspondence with theoretical results over all values of load pf.

A typical transient response is shown in Fig. 8 for a step application of a single-phase load using the 2-control loop system. As pointed out earlier, no special effort was made to improve the speed of response, overshoot, etc., since this was not the primary objective of the investigation. The observed speed of response was considered satisfactory, however, for most power applications.

Conclusions

Three-phase to single-phase power conversion can be accomplished without the use of large rotating equipment on limited power capacity systems. This conversion is accomplished with an attendant improvement of pf to 0.866.

It is possible that similar conversion from single-phase to 3-phase power may be effected, but the error-sensing and control loop operation will be considerably more complex than the simple one outlined.

References

1. Three Phase Distribution of Electric Welding and Furnace Loads, A. L. Morris. *Journal*, Institution of Electrical Engineers, London, England, vol. 93, pt. II, 1946, pp. 341–52.

2. A Unit for Balancing the Voltages of Three-Phase Aircraft Inverters Supplying Large Single-Phase Loads, Sydney Wimple. *AIEE Transactions*, vol. 63, July 1944, pp. 532–33.

A General Technique for Approximating Transient Response from Frequency Response Asymptotes

GEORGE A. BIERNSON
ASSOCIATE MEMBER AIEE

ALTHOUGH feedback control design is generally performed in terms of frequency response, dynamic behavior is best expressed in terms of transient response. Consequently it is very desirable to be able to relate transient and frequency response in a simple manner. This paper presents, as a means for doing this, a reliable approximation procedure, which enables one to relate by inspection the characteristics of the frequency response to the corresponding characteristics of the transient responses.

The term transient response includes, along with the step response, the responses to steps of derivatives and to steps of integrals of the input. As reference 1 shows, there is a set of transient terms which define the time response of a given system to any input. If the major characteristics of each transient of this set are known with reasonable accuracy, one can define the time response of the system to any input with reasonable engineering accuracy. Thus, it is important that one be able to estimate from frequency-response data the major characteristics of not only the step response but also the ramp response, the response to a step of acceleration, the impulse response, etc.

Reference 2 presents by intuitive reasoning a number of rules for approximating the shapes of the transient responses by inspecting frequency-response characteristics. However, it is difficult to apply these rules adequately without a more rigorous justification. It is the purpose of this paper to present a mathematical basis for justifying the approximations of reference 2 and to provide a basis for further approximations.

Paper 56-640, recommended by the AIEE Feedback Control Systems Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Summer and Pacific General Meeting, San Francisco, Calif., June 25–29, 1956. Manuscript submitted October 14, 1955; made available for printing April 17, 1956.

GEORGE A. BIERNSON is with Sylvania Electric Products, Inc., Waltham, Mass.

This research was supported by the Department of the Air Force under Air Force Contract *AF33(616)-2038* at the Servomechanisms Laboratory in the Department of Electrical Engineering at Massachusetts Institute of Technology.

The techniques presented in this paper are fundamentally based on reference 3, which shows how the closed-loop poles of a feedback control system are related to the open-loop frequency response. By expressing the computational procedures of reference 3 in terms of asymptotic magnitude plots, a method for approximating the values of the poles by inspection is derived. With the values of the poles known, a means of directly forming asymptotic plots of the closed-loop functions is developed. Finally the paper shows how the forms of the transient responses can be estimated directly from the values of the closed-loop asymptote plots at the various closed-loop poles.

When the approximation procedures presented in this paper and in reference 2 are combined with the method of reference 1, a direct procedure is provided for designing a system for a prescribed response to any given input. By means of reference 1, the required characteristics of the response to the given input can be expressed in terms of major characteristics of the various transient responses, and by the approximations of reference 2 and this paper, these transient response characteristics can in turn be related directly to the required characteristics of the open-loop frequency response.

Terminology

The terminology and symbolism to be used is illustrated in Fig. 1. The time-varying signals are represented by lower case letters and their transforms by upper case. Generally, signals are represented by the letter x or X with a subscript, and transfer functions by the letter G with a double subscript. For example, Fig. 1(A) shows that the transfer function of an element relating an input X_i to an output X_o is designated as G_{io}.

For a feedback control loop, three basic variables are defined, and these are the three variables existing about the point where the loop input is added into the loop. The signals, which are illustrated in Fig. 1(B), are

X_i = loop input
X_r = loop return signal
$X_e = X_i - X_r$ = loop error signal or loop difference signal

There are three main transfer functions relating these three variables

$$G = \frac{X_r}{X_i} = \text{loop transfer function}$$

$$G_{ir} = \frac{X_r}{X_i} = \text{return transfer function}$$

$$G_{ie} = \frac{X_e}{X_i} = \text{error transfer function}$$

According to the double-subscript notation, the loop transfer function could also be expressed as G_{ro} but for simplicity it is merely represented as G. It is often called the "open-loop" transfer function, and G_{ir} and G_{ie} may be called "closed-loop" transfer functions.

In practice, of course, one is often interested in determining the transfer function in a feedback control loop between the loop input X_i and a loop output X_o, as shown in Fig. 1(C). However, since the fundamental feedback control relationships are in terms of the basic loop variables X_i, X_e and X_r, it is best to redraw Fig. 1(C) in the forms of Fig. 1(D), so that X_o is related by a simple cascade element to either X_e or X_r. Thus the relationship between X_i and X_o should be considered to be composed of two parts, one defining the action of the feedback-control loop and represented by either G_{ie} or G_{ir}, and the other representing the effect of a cascade element. Consequently, a complete study of the characteristics of a feedback control loop may be made by considering only the transfer functions, G, G_{ie}, and G_{ir}.

Instead of using the well-known logarithmic unit, the decibel, which has the disadvantage of being defined basically in terms of a power ratio, the paper uses the new and more general logarithmic unit, the decilog(dg). The value of any ratio A in dg is

$$A \text{ in dg} = 10 \log_{10} A \qquad (1)$$

Thus, 10 dg represents a factor of 10.

Estimation of Values of Closed-Loop Poles by Inspection

In reference 3, a numerical-reiteration process is developed for computing the closed-loop poles of a feedback control system from the loop transfer function G. By means of approximations, this process can be simplified so that the closed-loop poles may be estimated by inspection from an asymptotic magnitude plot of G.

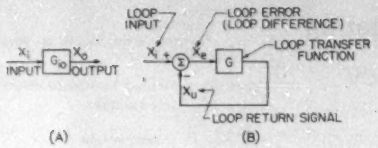
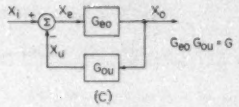


Fig. 1 (left). Feedback - loop terminology



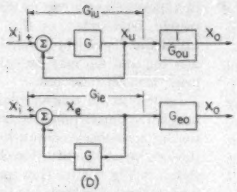


Fig. 3. Fictitious function F(s) for illustrative system

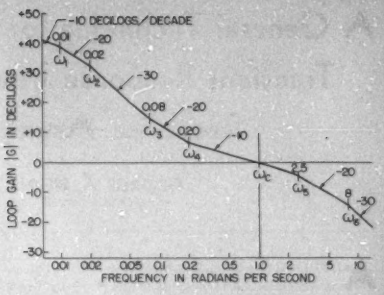
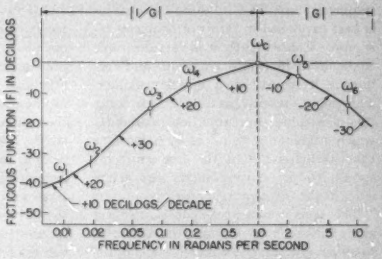


Fig. 2 (right). Loop transfer function for illustrative system



Rough Estimation from Asymptotes

Analysis

As shown in reference 3, the closed-loop poles of a feedback control loop with the loop transfer function G are approximately equal to:

1. The zeros of G below ω_c in frequency.
2. The poles of G above ω_c in frequency.
3. Plus the closed-loop pole $-\omega_c$.

where ω_c represents the gain-crossover frequency (the frequency where the loop gain $|G|$ is unity). Since this is only an approximation, it is also adequate to consider ω_c as the asymptote crossover frequency, which is the frequency where the -10 decilog-per-decade asymptote of G near gain crossover crosses unity gain.

A more unified statement of the values of the closed-loop poles may be obtained by defining a fictitious transfer function $F(s)$ which is equal to $1/G(s)$ for values of $|s|$ below ω_c, and is equal to $G(s)$ for values of $|s|$ above ω_c. In terms of $F(s)$ the approximate poles are, then:

1. The poles of F.
2. Plus the closed-loop pole $-\omega_c$.

The function $F(s)$ cannot be represented as a function of s for all values of s: it is merely an artifice which is convenient to employ in graphical calculations.

As an illustration, consider the loop transfer function shown in Fig. 2.

$$G=\frac{\omega_c(s+\omega_2)(s+\omega_4)}{s(s+\omega_1)(s+\omega_3)[(s/\omega_5)+1][(s/\omega_6)+1]} \tag{2}$$

where ω_c is the asymptote crossover frequency, set at 1.00 radian per second. In terms of the numerical values given in the figure for the break frequencies

$$G=\frac{(1.00)(s+0.08)(s+0.2)}{s(s+0.01)(s+0.02)[(s/2.5)+1]\times[(s/8)+1]} \tag{3}$$

Fig. 3 gives an asymptotic magnitude plot of the fictitious function F corresponding to the loop transfer-function G in Fig. 1. The approximate closed-loop poles are indicated in Fig. 3 by circles at the corresponding break frequencies, which represent the magnitudes of the approximate poles.

Fig. 3 shows that approximate closed-loop poles occur wherever the slope of the magnitude asymptote of F increases negatively. Except at the frequency ω_c, (which is not the break frequency for a pole of F but rather is the frequency where F is discontinuous) a first-order closed-loop pole occurs wherever the slope of F increases negatively by 10 decilogs per decade, a second-order closed-loop pole occurs wherever the slope increases negatively by 20 decilogs per decade, etc.

In reference 3, equations 38 and 39 give the following as first approximations (in the numerical-reiteration process) to the shift of an actual closed-loop pole from a single-order approximate pole

$$\delta_a \approx -[(s-s_a)/G(s)]_{s=s_a} \text{ for } |s_a|<\omega_c \text{ and } s\approx s_a \tag{4}$$

$$\delta_b \approx -[(s-s_b)G(s)]_{s=s_b} \text{ for } |s_b|>\omega_c \text{ and } s\approx s_b \tag{5}$$

In terms of the fictitious function F, equations 3 and 4 may be represented as one equation.

$$\delta_b \approx -[(s-s_b)F(s)]_{s=s_b} \text{ for } s\approx s_b \tag{6}$$

where s_b is now a pole of F and hence an approximate closed-loop pole, and δ_b is the shift of the exact closed-loop pole from this pole of F. The magnitude of this shift in equation 6 is

$$|\delta_b| \approx |(s-s_b)F(s)|_{s=s_b} \tag{7}$$

Since the factor $(s-s_b)$ cancels out the pole of $F(s)$ at $s=s_b$, the function $(s-s_b)F$ has neither a pole nor a zero at $s=s_b$. If there are no other poles or zeros of F near s_b, the magnitude of this function $(s-s_b)F$ can be approximated by its asymptote

$$|(s-s_b)F(s)|_{s=s_b} \approx \text{asm}\,|(s-s_b)F(s)|_{s=s_b} = |s_b|\,\text{asm}\,|F(s_b)| \tag{8}$$

where asm symbolizes the asymptote and asm $|F(s_b)|$ symbolizes the magnitude asymptote of $F(s)$ evaluated at $s=s_b$. Substitution of equation 7 into equation 6 gives

$$|\delta_b/s_b| \approx \text{asm}\,|F(s_b)| \tag{9}$$

Table I. Computation of Closed-Loop Poles from Asymptote

Pole of F (s_b) (1)	Asm $F(s_b)$ Dg (2)	Asm $F(s_b)$ Ratio (3)	Approximation $\lvert\delta_b/s_b\rvert$ (4)	Value of s_b (5)	Approximation $\lvert\delta_b\rvert$ (6)	Direction of Shift (7)	Closed-Loop Pole s_b' Approximation (8)	Closed-Loop Pole s_b' Exact (9)
$-\omega_3$	−15	0.0316	0.0316	−0.08	0.0025	decrease	−0.0775	−0.0776
$-\omega_4$	− 7	0.20	0.20	−0.2	0.040	increase	−0.24	−0.2885
$-\omega_5$	− 4	0.40	0.40	−2.5	1.0	decrease	−1.5	$\omega_n = 1.308$, $\zeta = 0.67$
$-\omega_6$	−14	0.040	0.040	−8.0	0.32	increase	−8.32	−8.392

Thus, the magnitude of the per-unit shift of a closed-loop pole from a single-order open-loop pole of F is roughly equal to the magnitude asymptote of F at the open-loop pole.

If there were a double-order pole of F at s_b, it would form two closed-loop poles shifted in opposite directions. The magnitude of each shift would be roughly the following:

$$|\delta_b/s_b| \approx \sqrt{\text{asm}\,|F(s_b)|} \qquad (10)$$

To compute the closed-loop poles it is necessary to consider the directions of the shifts. As shown in Appendix II of reference 3, the closed-loop poles shift from the poles of F along the loci where the phase is 180 degrees. A single-order pole of F has one such locus emerging; a double-order pole has two; a triple-order pole, three, etc. Along the negative real axis, the phase is 180 degrees wherever the slope of the magnitude asymptote is an odd multiple of ±10 decilogs per decade, and zero where it is an even multiple. Consequently, from a single-order real pole of F the closed-loop pole shifts along the negative real axis into the region where the magnitude asymptote slope is an odd multiple of 10 decilogs per decade.

Example

In Table I, equation 9 is employed to calculate the closed-loop poles for the illustrative system of Fig. 2. As shown in Fig. 3, the system has five approximate closed-loop poles: $-\omega_3$, $-\omega_4$, $-\omega_5$, $-\omega_6$, and $-\omega_c$. The decilog values of the magnitude asymptote of F at the frequencies ω_3, ω_4, ω_5, and ω_6, read from Fig. 3, are tabulated in column 2 of Table I, and the ratios corresponding to these decilog values are tabulated in column 3. By equation 9, these ratios are approximately equal to the per-unit shift $|\delta_b/s_b|$ as indicated in column 4. In column 5, the values of the approximate poles are listed. Multiplying the magnitudes in column 5 by the corresponding approximate per-unit shifts in column 4 give the magnitudes of the approximate shifts in column 6.

Since all of the poles of F are real and of single order, the closed-loop poles must shift from the poles of F along the negative real axis into frequency regions where the asymptote slope is an odd multiple of 10 decilogs per decade. The frequency regions with these slopes are as follows: $0<|s|<\omega_1$; $\omega_2<|s|<\omega_3$; $\omega_4<|s|<\omega_5$; $\omega_6<|s|<\infty$. Therefore the shifts from the poles at ω_3 and ω_6 are toward decreasing frequency, while the shifts from the poles at ω_4 and ω_5 are toward increasing frequency, as is indicated in column 7 of Table I. Depending upon the direction of shift, the magnitudes of the individual shifts in column 6 are either added to or subtracted from the values of the poles of F in column 5 to obtain the calculated closed-loop poles listed in column 8. For example, the closed-loop pole $-\omega_3'$, which is shifted from $-\omega_3$, is calculated as

$$-\omega_3' = -(\omega_3 - |\delta_3'|) = -(0.08 - 0.0025) = -0.0775 \qquad (11)$$

The exact values of the closed-loop poles are listed in column 9 for comparison. For the closed-loop poles shifted from $-\omega_3$ and $-\omega_4$, the calculated values in column 8 agree very well with the exact values in column 9. For the other two poles the agreement is much poorer because the magnitude asymptote of F at ω_4 and ω_5 is quite large, resulting in a large per-unit shift.

The closed-loop pole shifted from $-\omega_5$ is combined with the pole shifted from $-\omega_c$ to form an underdamped pair, with the damping ratio ζ and natural frequency ω_n given in column 9. The approximate value −1.5 given in column 8, for the pole shifted from $-\omega_5$, leads one to suspect this, because −1.5 is so close to $-\omega_c$.

Effect of Deviation from Asymptote

Analysis

For a better approximation of the closed-loop poles, the right-hand expression of equation 7 may be evaluated exactly by multiplying the asymptote of $|F(s_b)|$ by the deviations from the asymptote at s_b of the individual poles and zeros of $(s-s_b)F$. Fig. 4 gives a family of curves, obtained from reference 3, for computing these deviations. If s_b is real, the index line of Fig. 4, at $|s/s_b|=1$, is placed at the frequency $|s_b|$, and the deviation in decilogs at each pole and zero of F, except the pole at s_b, is read from the curve for the ζ of that pole, or zero. The deviation due to the pole of F at s_b (which is infinite) is not considered, because $(s-s_b)F$ does not contain that pole. When there are no other poles or zeros of F near s_b in magnitude, these deviations are small, and hence the asymptotic approximation of $|(s-s_b)F|$ is very good. The closer a pole or zero is to s_b, the larger its deviation.

The deviations given in Fig. 4 are equal to

$$\text{dev (deviation)} = \frac{|s-s_1|}{\text{asm}\,|s-s_1|_{s=s_b}} = \frac{|s_2-s_1|}{\text{asm}\,|s_2-s_1|} = \frac{|\Delta_s|}{|s_b|} \qquad (12)$$

where s_1 and s_2 are the two complex frequencies concerned. The difference between these frequencies is Δ_s, and s_b is either s_1 or s_2, depending upon which is the

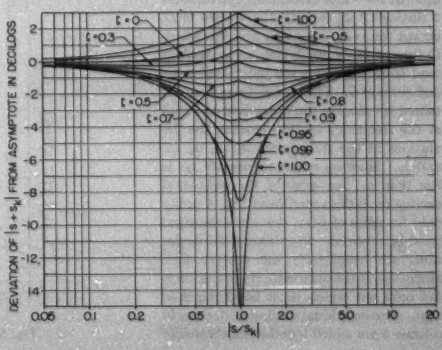


Fig. 4. Magnitude deviation for zero at $s = s_b$

Table II. Computation of Deviation from Asymptote at ω_4

Pole or Zero of G			Individual Deviation, Dg		Total Deviation	
Symbol (1)	Value (2)	Type (3)	From Fig. 3 (4)	Correct Sign (5)	Dg (6)	Ratio (7)
$-\omega_1$	−0.01	pole	−0.23	−0.23		
$-\omega_2$	−0.02	pole	−0.46	−0.46		
$-\omega_3$	−0.08	zero	−2.22	+2.22	1.06	1.27
$-\omega_4$	−2.5	pole	−0.36	−0.36		
$-\omega_5$	−8.0	pole	−0.11	−0.11		

greater in magnitude. The deviation thus is the ratio of the distance between the two values of s, divided by the distance of the further of the two from the origin.

The decilog deviations in Fig. 4 are either added to or substracted from the decilog value of the asymptote of F, depending upon the relative position of the individual poles or zeros. To determine whether to add or substract, it is easier to consider G rather than F. In terms of F, the deviation being computed is symbolized as

$$\text{dev}\,|(s-s_b)F|_{s=s_b} \tag{13}$$

In terms of G, this expression has two forms, depending on whether $|s_b|$ is greater than or less than ω_c, which are

$$\text{dev}\,|(s-s_b)G|_{s=s_b}, \text{ for } |s_b|>\omega_c \tag{14}$$

$$\text{dev}\,|(s-s_b)/G|_{s=s_b}, \text{ for } |s_b|<\omega_c \tag{15}$$

It can be shown readily from these expressions that the rules for the proper sign of the individual decilog deviations may be expressed as follows.

For $|s_b|$ greater than ω_c, the deviation curves of Fig. 4 as shown apply for zeros of G; for poles of G invert the curves.

For $|s_b|$ less than ω_c, the deviation curves of Fig. 4 as shown apply for poles of G; for zeros of G invert the curves.

Example

In Table II, the deviation for $|(s+\omega_4)F|$ at $s=-\omega_4$ is calculated. In columns 1 through 3 all the poles and zeros of G, except for $-\omega_4$, are listed and described. Column 4 gives the deviations in decilogs read from Fig. 3 for the individual poles and zeros of G. These deviations were obtained by placing the index line of Fig. 4 at the frequency ω_4, and reading the deviations at the frequencies ω_1, ω_2, ω_3, ω_5 and ω_8. Since ω_4, which represents $|s_b|$, is less than ω_c, the signs of the individual decilog deviations read from Fig. 4 apply only for poles of $G(s)$; for zeros the signs must be reversed. In column 5 the deviations are listed with the correct sign. Since $-\omega_3$ is the only zero, its deviation is the only one which must be changed in sign. The individual decilog deviations in column 5 are added together algebraically to obtain the total decilog deviation in column 6. In column 7 this decilog deviation is converted to the ratio value 1.27.

With this deviation, the exact value $|(s+\omega_4)F(s)|$ for $s=-\omega_4$, can be calculated as follows:

$$|(s+\omega_4)F|_{s=-\omega_4}=\omega_4 \text{ asm}\,|F(\omega_4)|\times \text{dev}\,|(s+\omega_4)F|_{s=-\omega_4} \tag{16}$$

From Table I, ω_4 is equal to 0.20 and the magnitude asymptote of F at ω_4 is equal to 0.20; from Table II the deviation of F at ω_4 is equal to 1.27. Hence equation 16 is equal to

$$|(s+\omega_4)F|_{s=-\omega_4}=(0.20)(0.20)(1.27)=0.51 \tag{17}$$

By equation 7, this is approximately equal to $|\delta_4|$. Therefore the closed-loop pole $-\omega_4'$ is approximately

$$-\omega_4' \approx -(\omega_4+|\delta_4|)=-(0.20+0.51) =-0.251 \tag{18}$$

This value is still significantly less than the exact value for $-\omega_4'$, which is -0.2885, but it is closer than the first approximation -0.24 calculated in Table I.

Exact Calculation of Poles

Analysis

The expression for $|\delta_b|$ in equation 6 is approximate because the function $|(s-s_b)F|$ is evaluated at the approximate closed-loop pole, $s=s_b$, rather than at the exact closed-loop pole. Now, the exact closed-loop pole is equal to

$$s_b'=s_b+\delta_b \tag{19}$$

Hence, the exact expression for $|\delta_b'|$ is

$$|\delta_b|=|(s-s_b)F|_{s=s_b+\delta_b} \tag{20}$$

Of course, equation 20 cannot be solved explicitly, because to obtain the solution requires a knowledge of the exact closed-loop pole $(s+\delta_b)$. However, as shown in reference 3, part III, δ_b may be computed by a numerical reiteration process in which the best approximation for the closed-loop pole so far obtained in substituted for s in the expression $(s-s_b)F$.

Assume that the approximate pole s_b is real, and designated as $-\omega_b$. If the first step in the reiteration process gave a value δ_{b1} for the shift, a closer approximation δ_{b2} can be obtained from

$$|\delta_{b1}|=|(s-s_b)F(s)|_{s=-\omega_b+\delta_1=-\omega_z} \tag{21}$$

where $-\omega_z$ is defined as $-(\omega_b-\delta_{b1})$, the approximation to the pole evaluated by the first step in the reiteration process. Equation 21 may be expressed in terms of the magnitude asymptote of F evaluated at $s=-\omega_z$, as follows:

$$\delta_{b2}=\omega_{z+}\times\text{asm}\,|F(\omega_z)|\times \text{dev}\,|(s+\omega_b)F|_{s=-\omega_z} \tag{22}$$

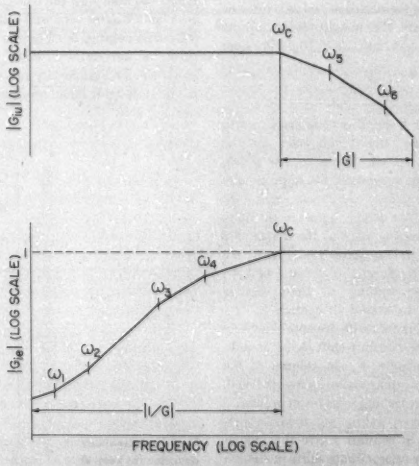


Fig. 5. Rough approximation of closed-loop magnitude plots

Table III. List of Break Frequencies

Function	General Expression	Break Frequencies
$G_{iu}=\frac{X_u}{X_i}=G$	$\frac{A}{B}$	$0,\ \omega_1,\ \omega_2 \mid \omega_3,\ \omega_4$
$G_{iu}=\frac{X_u}{X_i}=\frac{G}{1+G}$	$\frac{A}{(A+B)}$	$\omega_1'\ \ \omega_2'\ \mid\ (\omega_3')^2\ \ \omega_4'$
$G_{ie}=\frac{X_e}{X_i}=\frac{1}{1+G}$	$\frac{B}{(A+B)}$	$0,\ \omega_1,\ \omega_2 \mid \omega_3,\ \omega_4$
$G_{ue}=\frac{X_e}{X_u}=\frac{1}{G}$	$\frac{B}{A}$	$0,\ \omega_1,\ \omega_2 \mid \omega_3,\ \omega_4$

where $\omega_{b\varepsilon+}$ is the asymptote of $|s-s_b|$ evaluated at $s=-\omega_z$, and hence is equal to either ω_b or ω_z, whichever is greater.

Example

Equation 22 may be applied as follows to give a closer approximation of the pole shifted from $-\omega_4$. The best approximation thus far obtained for the pole shifted from $-\omega_4$ was given in equation 18, as $-\omega_4'=0.251$. In equation 22, ω_z represents ω_4', ω_b represents ω_4 which is -0.200, and ω_{bz+} is equal to ω_z, since ω_z is greater than ω_b. This gives for the second approximation of $|\delta_4|$

$$|\delta_{42}|=\omega_4'\times \text{asm}\,|F(\omega_4')|\times \text{dev}\,|(s+\omega_4)F|_{s=-\omega_z} \quad (23)$$

$$=(0.251)\text{asm}\,|F(0.251)|\times \text{dev}\,|(s+\omega_4)F|_{s=-0.251} \quad (24)$$

From Fig. 3, the asymptote of F at the frequency 0.251 is

$$\text{asm}\,|F(0.251)|=-6\ d_E=0.25 \quad (25)$$

As a reasonable approximation, assume that the deviation of $|(s+\omega_4)F|$ has the same value at $-\omega_4'$ that it had at $-\omega_4$, which was 1.27. Substituting this deviation and equation 25 into equation 24 gives for the second approximation of $|\delta_4|$

$$|\delta_{42}|=(0.251)(0.25)(1.27)=0.080 \quad (26)$$

Since the shift is toward infinite frequency, the value of the closed-loop pole is

$$-\omega_4'\approx-(\omega_4+|\delta_{42}|)=-(0.200+0.080)=-0.280 \quad (27)$$

The exact value of $-\omega_4'=-0.288$; the approximate value previously calculated in equation 21 was -0.251. Although this reiteration process may be repeated, two tries should be sufficient for the purpose of this method, either to yield a satisfactory answer or to show that the reiteration process will not converge rapidly enough to make the method practical.

Forming Magnitude Asymptote Plots of Closed-Loop Transfer Functions

METHOD FOR FORMING PLOTS

Once the closed-loop poles have been computed, asymptotic magnitude plots of the closed-loop transfer functions G_{iu} and G_{ie} can be developed quite simply from an asymptotic magnitude plot of F. As will be shown, asymptotic magnitude plots of the closed-loop transfer functions are quite useful for estimating the transient response.

If the loop transfer function G is expressed as follows as a ratio of two polynomials in s

$$G=\frac{X_u}{X_s}=\frac{A(s)}{B(s)} \quad (28)$$

the closed-loop transfer functions can be expressed as

$$G_{iu}=\frac{X_u}{X_i}=\frac{G}{1+G}=\frac{A(s)}{A(s)+B(s)} \quad (29)$$

$$G_{ie}=\frac{X_e}{X_i}=\frac{1}{1+G}=\frac{B(s)}{A(s)+B(s)} \quad (30)$$

At low frequencies (i.e., for $|s|$ much less than ω_c) G_{iu} is approximately unity and G_{ie} is approximately equal to $(1/G)$; while at high frequencies (i.e., for $|s|$ much greater than ω_c) G_{iu} is approximately equal to G and G_{ie} is approximately unity. Consequently, the closed-loop functions may be approximated roughly

$$G_{iu}\approx 1,\ G_{ie}\approx(1/G),\ \text{for}\ |s|<\omega_c \quad (31)$$

$$G_{iu}\approx G,\ G_{ie}\approx 1,\ \text{for}\ |s|>\omega_c \quad (32)$$

Fig. 5 gives asymptotic magnitude plots of these approximations. Note that the plot of F coincides with the approximate plot of G_{iu} below ω_c, and with that of G_{ie} above ω_c. These approximations are poor near ω_c, but are increasingly exact as $|s|$ approaches zero or infinity.

To form exact asymptotic magnitude plots of these closed-loop functions, one must know the break frequencies (magnitudes) of most of the closed-loop poles, although a knowledge of all of them is not necessary. Assume that the three closed-loop poles, $-\omega_3'$, $-\omega_4'$, and $-\omega_6'$, for the example are computed with sufficient accuracy. The loop transfer function G shown in Fig. 2 is equal to

$$G=\frac{\omega_c(s+\omega_2)(s+\omega_4)\omega_5\omega_6}{s(s+\omega_1)(s+\omega_3)(s+\omega_5)(s+\omega_6)} \quad (33)$$

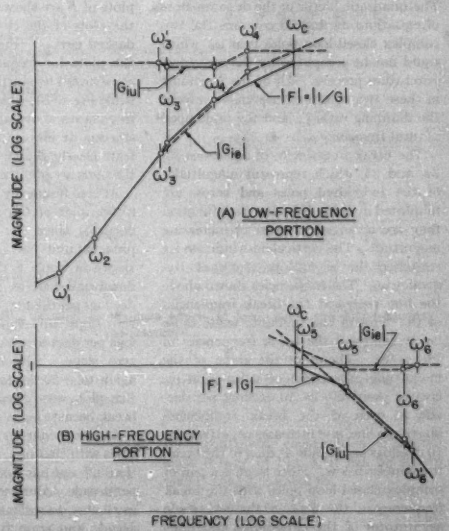


Fig. 6. Formation of closed-loop magnitude plots

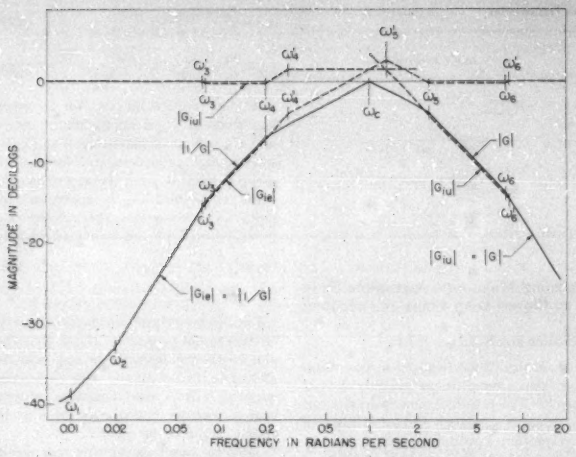


Fig. 7. Frequency plots of closed-loop functions drawn to scale

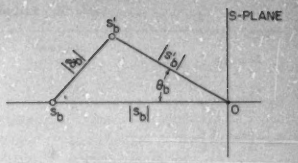


Fig. 8. Shift of complex closed-loop pole from pole of F(s)

By equations 29 and 30, the closed-loop transfer functions are

$$G_{1u} = \frac{\omega_c(s+\omega_3)(s+\omega_4)\omega_5'\omega_6}{(s+\omega_3')(s+\omega_4')\times (s^2+2\zeta_5'\omega_5's+\omega_5'^2)(s+\omega_6')} \quad (34)$$

$$G_{1\ell} = \frac{s(s+\omega_1)(s+\omega_2)(s+\omega_3)(s+\omega_4)}{(s+\omega_3')(s+\omega_4')\times (s^2+2\zeta_5'\omega_5's+\omega_5'^2)(s+\omega_6')} \quad (35)$$

The quadratic factor in the denominators of equations 34 and 35 contains the two complex closed-loop poles near ω_c, which could not be computed by the numerical reiteration process. All of the constants in these two equations are known except the damping ratio ζ_5', and the undamped natural frequency ω_5'.

The break frequencies of equations 33, 34, and 35, which represent magnitudes of the individual poles and zeros, are tabulated in Table III. For each function they are arranged in order of increasing magnitude. The vertical lines indicate for reference the asymptote crossover frequency ω_c. The frequencies shown above the line represent the break frequencies of the zeros of the function, while those below represent the break frequencies of the poles. Therefore the slope of the magnitude asymptote of a function increases positively by 10 decilogs per decade at each of the break frequencies above the line, and increases negatively by 10 decilogs per decade at each of the break frequencies below. Since there is a pair of complex closed-loop poles with the break frequency ω_5' the slopes of the magnitude asymptotes of the closed-loop functions increase negatively by 20 decilogs per decade at ω_5', which is indicated in the table by showing ω_5' as squared. This symbol ω_5' is also placed in parentheses to indicate that the value for ω_5' is unknown.

With the help of Table III, the asymptote magnitude plots of the closed-loop functions are developed in Fig. 6 from the asymptotic magnitude plots of F. The plots of F are shown as solid curves, and the plots of the closed-loop functions as dashed curves. Fig. 6(A), which shows the plots for frequencies below ω_c, was constructed by starting at zero frequency; while Fig. 6(B), which shows the plots for frequencies above ω_c, was constructed by starting at infinite frequency. To illustrate clearly the method of construction, the plots are not drawn to scale.

At zero frequency, the asymptotic magnitude plot of G_{1u} starts at unity (zero decilogs) since (by equation 31) G_{1u} approaches unity as $|s|$ approaches zero. As shown in Table I, the plot of G_{1u} breaks downward first at ω_2' to a slope of -10 decilogs per decade; then up again at ω_3 to zero slope, up again at ω_4 to $+10$ decilogs per decade, and down again at ω_4' to zero slope. Although it breaks down again to -20 decilogs per decade at ω_5', the plot was discontinued before this break because ω_5' is unknown.

At zero frequency, the plot of $G_{1\ell}$ coincides with that of $1/G$, as shown by equation 32, and has a slope of $+10$ decilogs per decade. Along with $1/G$, it breaks upward at ω_1 to a slope of $+20$ decilogs per decade, and again at ω_2 to a slope of $+30$ decilogs per decade. At ω_3' the plot of $G_{1\ell}$ departs from that of $1/G$, breaking downward, to $+20$ decilogs per decade; and at ω_4' it breaks downward to $+10$ decilogs per decade. At ω_5' it breaks downward again to a slope of -10 decilogs per decade, but since ω_5' is unknown, the plot was discontinued after ω_4'. Note that the plot of $G_{1\ell}$ intersects that of G_{1u} at ω_c. This must always happen, as will be shown.

In Fig. 6(B) these plots are constructed for frequencies above ω_c, by starting at infinite frequency and progressing toward zero frequency. At infinite frequency the magnitude asymptote of G_{1u} coincides with that of G. As it progresses toward zero frequency, it breaks at ω_6' as shown. At the frequency ω_5', as yet unknown, it should break again to follow along the asymptotic plot constructed in Fig. 6(A) for frequencies below ω_c. Therefore, if these two plots of G_{1u} in Figs. 6(A) and 6(B) are superimposed, the frequency at which they intersect must be the unknown frequency ω_5'. Thus ω_5', the break frequency of the complex pair of closed-loop poles, can be computed by a simple graphical construction, when the break frequencies of the other closed-loop poles are known.

In Fig. 7, the asymptotic magnitude plots of G_{1u} and $G_{1\ell}$ are constructed to scale. The frequency at which the low-frequency and high-frequency portions of G_{1u} cross is shown in Fig. 7 to be $\omega_5' = 1.30$. Considering the accuracy of the graphical construction, this is in very good agreement with the exact value of ω_5', 1.308, which was given in column 9 of Table I.

The frequency ω_5' can also be computed from the plots of $G_{1\ell}$. As shown in Fig. 6(B), the magnitude asymptote of $G_{1\ell}$ is unity at infinite frequency. As it progresses toward zero frequency, it breaks at ω_6', then back to zero slope at ω_5', and to a -10 decilog per decade slope at ω_3. If this plot in Fig. 6(B) is superimposed on the plot of $G_{1\ell}$ in Fig. 6(A), the frequency at which the asymptotic plots cross must also be ω_5'. In Fig. 7, these asymptotic

plots are drawn to scale and cross at the frequency $\omega_3' = 1.31$.

If the value of the break frequency ω_3 in the transfer function of G were increased sufficiently, there would no longer be a pair of complex closed-loop poles near ω_c. Instead, there would be two real closed-loop poles: one, $-\omega_c'$, shifted from $-\omega_c$; and the other, $-\omega_3'$, shifted from $-\omega_3$. The frequency ω_3' obtained from the graphical constructions shown in Fig. 7, then, would not represent the undamped natural frequency of a pair of complex poles; but it would represent the geometric mean of the two closed-loop break frequencies ω_c' and ω_3', and would be equidistant from these two frequencies on a log-frequency plot.

It is easier to construct the closed-loop asymptotic magnitude plots when one recognizes that, at any given frequency, the exact asymptotic plots of G_{iu} and G_{ie} must deviate from their approximate plots, given in Fig. 5, by the same amount. For frequencies below ω_c, the asymptotic plot of G_{ie}, is the same distance above or below the asymptotic plot of $1/G$ as the asymptotic plot of G_{iu} is above or below the zero decilog line, which is indicated by the following identity:

$$\frac{G_{ie}}{(1/G)} = \frac{X_e/X_i}{X_e/X_u} = \frac{X_u/X_i}{1} = \frac{G_{iu}}{1} \qquad (36)$$

The magnitude asymptotes of the functions in equation 36 must be in the same proportion, because equation 36 is an identity and hence must hold for all values of s. Similarly, for frequencies above ω_c, the asymptotic plot of G_{iu} is the same distance above or below the asymptotic plot of G as the asymptotic plot of G_{ie} is above or below the zero-decilog axis, which is indicated by the identity:

$$\frac{G_{iu}}{G} = \frac{X_u/X_i}{X_u/X_e} = \frac{X_e/X_i}{1} = \frac{G_{ie}}{1} \qquad (37)$$

At the frequency ω_c, the asymptotic magnitude plots of G_{iu} and G_{ie} must be equal, because the plot of G is at zero decilogs at ω_c, and the plot of G_{iu} is the same distance above that of G as the plot of G_{ie} is above the zero decilog line. Thus, as was stated previously, the asymptotic magnitude plots of the closed-loop functions must intersect at ω_c.

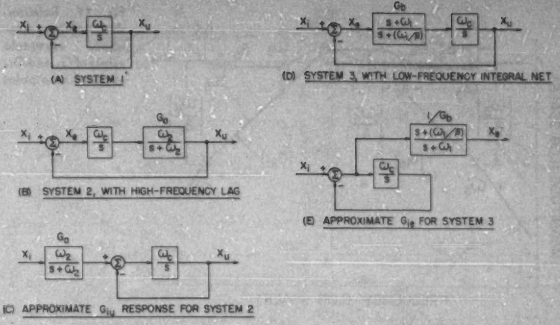

(A) SYSTEM 1
(B) SYSTEM 2, WITH HIGH-FREQUENCY LAG
(C) APPROXIMATE G_{iu} RESPONSE FOR SYSTEM 2
(D) SYSTEM 3, WITH LOW-FREQUENCY INTEGRAL NET
(E) APPROXIMATE G_{ie} FOR SYSTEM 3

Fig. 9. Block diagrams showing approximate effects of low-, mid-, and high-frequency characteristics

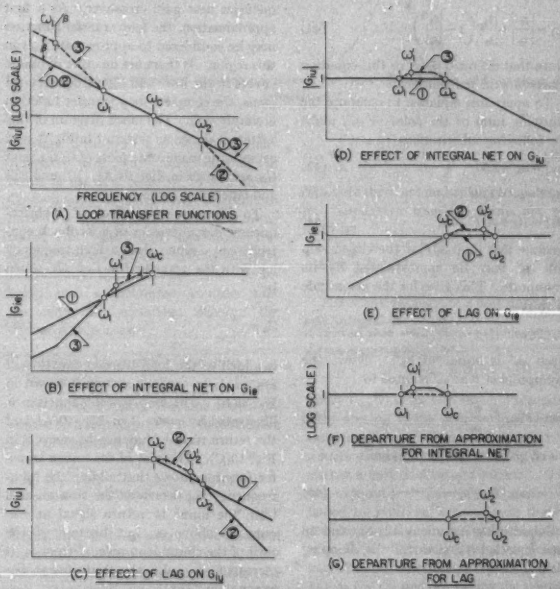

FREQUENCY (LOG SCALE)
(A) LOOP TRANSFER FUNCTIONS
(B) EFFECT OF INTEGRAL NET ON G_{ie}
(C) EFFECT OF LAG ON G_{iu}
(D) EFFECT OF INTEGRAL NET ON G_{iu}
(E) EFFECT OF LAG ON G_{ie}
(F) DEPARTURE FROM APPROXIMATION FOR INTEGRAL NET
(G) DEPARTURE FROM APPROXIMATION FOR LAG

Fig. 10. Frequency plots showing approximate effects of low-, mid-, and high-frequency characteristics

Computation of Damping Ratio of Complex Poles

Although the natural frequency for the poles at ω_3' was calculated from the plots in Fig. 7, the damping ratio of these poles was not obtained. However, once the natural frequency has been found, the damping ratio can be computed by again considering the magnitude of the shift.

By equation 24 the magnitude of the shift δ_b of a closed-loop pole s_b' from a pole s_b of F is

$$|\delta_b| = |(s - s_b)F|_{s=s_b'} \qquad (38)$$

The shift in the complex s-plane of a complex pole s_b' from a real pole s_b is illustrated in Fig. 8. The damping ratio ζ_b' of the complex pole s_b' is related to the angle θ_b, shown in the diagram by

$$\zeta_b' = \cos\theta_b \qquad (39)$$

By trigonometric construction on the triangle $0\text{-}s_b\text{-}s_b'$ of Fig. 8, it can be shown readily that ζ_b' is equal to

$$\zeta_b' = \frac{|s_b|^2 + |s_b'|^2 - |\delta_b|^2}{2|s_b|\,|s_b'|} \qquad (40)$$

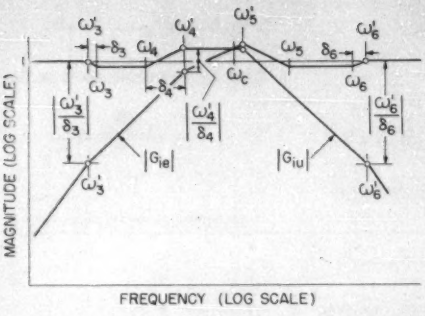


Fig. 11. Relationship between magnitude asymptote values of G_{ie} and G_{iu} at closed-loop poles

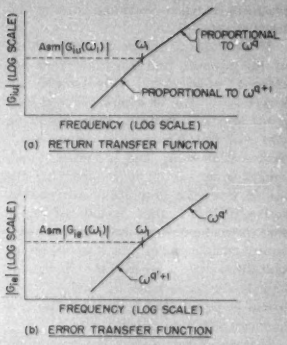


Fig. 12. Determination of transient coefficient sign from slope of magnitude asymptote

This equation holds, regardless of whether $|s_b|$ or $|s_b'|$ is greater. More conveniently

$$2\zeta_b = \left|\frac{s_b'}{s_b}\right| + \left|\frac{s_b}{s_b'}\right|\left(1-\left|\frac{\delta_b}{s_b}\right|^2\right) \tag{41}$$

Note that the ratio $|\delta_b/s_b|$ in this equation represents the per-unit shift.

To apply this equation to compute the damping ratio of the poles at ω_b', make the following substitutions

$$|s_b| = \omega_b = 2.50 \tag{42}$$

$$|s_b'| = \omega_b' = 1.30 \tag{43}$$

$$|\delta_b| = |(s+\omega_b)F|_{s=s_b'} \tag{44}$$

Assume that $|(s+\omega_b)F|$ at the closed-loop pole s_b' may be approximated by its asymptote. This gives for the magnitude of shift

$$|\delta_b| \cong \text{asm } |(s+\omega_b)F|_{s=s_b'} = \text{asm } |\omega_b F(\omega_b')|$$

since ω_b' is equal to $|s_b'|$. Now, the asymptote of F at ω_b' is equal to

$$\text{asm } |F(\omega_b')| = \frac{\omega_{b'}}{\omega_b'} \tag{45}$$

which gives for the per-unit shift

$$\left|\frac{\delta_b}{\omega_b}\right| = \text{asm } |F(\omega_b')| = \frac{\omega_c}{\omega_b'} = \frac{1}{1.30} \tag{46}$$

Substituting equations 42, 43, and 46 into equation 41 gives for the damping ratio

$$\zeta_b \cong 0.65 \tag{47}$$

which is in good agreement with the exact value 0.678.

Relative Effects on Response of Various Frequency Regions of G

A study of the asymptotic magnitude plots of the open- and closed-loop transfer functions of a feedback control loop can give insight into the relative affects of low-frequency, mid-frequency, and high-frequency portions of the loop transfer function.

The mid-frequency region of a feedback control loop corresponds to frequencies near gain crossover. As a first approximation, the loop transfer function may be considered to approximate ω_c/s in this region. If there are no other dynamic effects in the low- and high-frequency regions, the complete loop transfer function is simply ω_c/s. The block diagram of this system is shown as system 1 in Fig. 9, and asymptotic magnitude plots of G, G_{ie}, and G_{iu} are shown in Fig. 10(A), (B), and (C) and labelled as curve 1.

To illustrate the effect of high-frequency characteristics of a feedback control loop, assume that a high-frequency lag with the transfer function G_a, given by

$$G_a = \frac{\omega_2}{s+\omega_2} \tag{48}$$

is added to the loop transfer function of system 1, to produce system 2 shown in Fig. 9(B). The loop transfer function is illustrated by curve 2 in Fig. 10(A) and the return transfer function by curve 2 in Fig. 10(C). The plot of the return transfer function shows that adding the high-frequency lag decreases the transmission from the input to return signal at frequencies above ω_2, but because of the shift of the closed-loop pole ω_2' from ω_2, it increases, or peaks, the transmission between ω_c and ω_2.

If the peaking of G_{iu} between ω_c and ω_2, shown by the dashed curve in Fig. 10(C), is neglected, the resultant return transfer function for system 2 is the same as the transfer function in Fig. 9(C) between x_t and x_u. In Fig. 9(C) the input x_t is passed through the filter G_a and then to the original system 1. Thus, as a first approximation, to add a high-frequency lag G_a to a loop transfer function has the same effect on the return signal as passing the input signal through the lag G_a (which acts as a low-pass filter) before feeding it to the original loop. On the other hand, because the high-frequency lag adds phase lag to the loop, it produces peaking in the return transfer function, which may result in overshoot in the transient response, and possibly even instability.

To illustrate the effect of the low-frequency characteristics of a feedback control loop, assume a low-frequency integral network with the transfer function G_b, given by

$$G_b = \frac{s+\omega_1}{s+(\omega_1/\beta)} \tag{49}$$

is added to the loop transfer function of system 1, to produce system 3 shown in Fig. 9(D). The loop and error transfer functions are illustrated by curve 3 in Fig. 10(A) and (B). The plot of the error transfer function shows that adding the integral network decreases the error signal at frequencies below ω_1, but increases it between ω_1 and ω_e.

If the peaking of G_{ie} between ω_1 and ω_e is neglected, the resultant error transfer function for system 3 is the same as the transfer function in Fig. 9(E) between x_t and x_e. In Fig. 9(E) the error signal from the original system is passed through a filter with a transfer function equal to $1/G_b$ to produce the resultant error signal x_e. Thus, as a first approximation, to add a low-frequency integral network G_b to a loop transfer function has the same effect on the error signal as passing the original error signal through the network $1/G_b$. Since $1/G_b$ acts as a high-pass filter, with unity gain at high frequencies and attenuation below ω_1, the addition of the integral network G_b has the effect of attenuating, or filtering out, the frequency

components of the error signal below ω_1. On the other hand, since the integral network adds phase lag to the loop, it produces peaking in the closed-loop responses, resulting in a decrease in stability.

Adding a high-frequency lag also changes the error transfer function, just as adding a low-frequency integral network must also change the return transfer function. Fig. 10(D) compares the return transfer functions of systems 1 and 3, and Fig. 10(E) compares the error transfer functions of systems 1 and 2. Fig. 10(E) shows that the high-frequency lag adds peaking between ω_5 and ω_3 to both the return and error transfer functions, and Fig. 10(D) shows that the low-frequency integral network adds peaking between ω_1 and ω_6 to both transfer functions. Plots of the peaking in the return and error transfer functions are shown in Fig. 10(G) for the high-frequency lag, and Fig. 10(F) for the integral network. These plots of peaking, in a sense, define the departure of the actual system performance from the approximate performance illustrated by Fig. 9(C) and (E).

The purpose of this discussion has been to explain in a reasonably quantitative manner the main effects of dynamic factors in various frequency regions of the transfer function of a feedback control loop. That the addition of high-frequency lags has primarily the effect of filtering the input, and reducing loop stability, has generally been well known; but the main effects of adding integral networks has generally not been so well understood. The concept that an integral network acts like a high-pass filter on the error signal, attenuating or "integrating down" low-frequency components, can be very helpful in the design of feedback control systems.

Estimating Transient Responses

As shown in reference 1, the response of a system to any input can be calculated quite simply and accurately if one has reasonably accurate knowledge of a set of transient responses of the system. These are the responses to a step input, to steps of the first few derivatives and to steps of the first few integrals. Each response contains a transient portion and often a steady-state portion defined by steady-state (or "error") coefficients. An important consideration is that each response need not be known very accurately provided the complete set of responses is known.

Thus, if reasonably accurate plots of the set of transient responses are given, the response of the system is completely characterized with reasonable accuracy in terms of its time response to any input. Consequently, it is extremely desirable to have a simple approximation technique which can yield from asymptotic frequency response data reasonably accurate plots of these transient responses. To develop such a technique, consider first the important equations defining the transient responses.

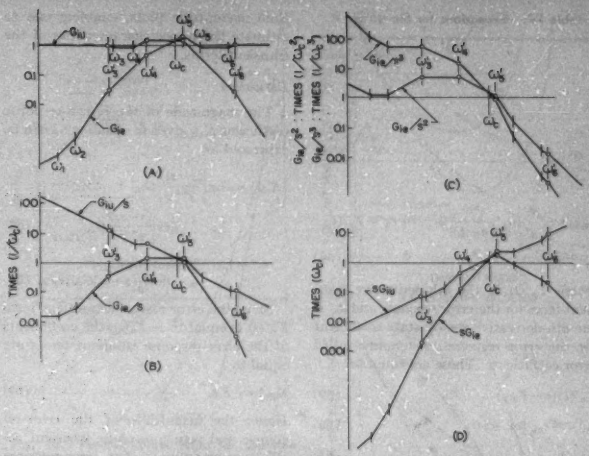


Fig. 13. Plots for estimating transient coefficients for steps of derivatives and for steps of integrals

Equations Defining Transient Responses

Reference 1 shows that the return response to a unit step of the nth derivative may be expressed as

$$x_n(t) = c_0 \frac{t^n}{n!} + c_1 \frac{t^{n-1}}{(n-1)!} \pm \ldots + c_n + T_n(t) \tag{50}$$

where $T_n(t)$ is the nth-derivative transient term for the return response, and c_n is the nth-derivative steady-state coefficient for the return response. The steady-state coefficients c_n are equal to

$$c_n = \frac{1}{n!}\left[\frac{d^n}{ds^n} G_{1u}\right]_{s=0} \tag{51}$$

If there are no multiple closed-loop poles, the transient term may be expressed as

$$T_n(t) = \sum_{l=1}^{L} K_{ln} e^{s_l t} \tag{52}$$

where s_l represents any one of the L closed-loop poles, and K_{ln} is the corresponding coefficient, equal to

$$K_{ln} = \left[\frac{(s-s_l)}{s^{n+1}} G_{1u}\right]_{s=s_l} \tag{53}$$

Whether there are multiple-order poles or not, successive transient terms are related by

$$T_n(t) = \frac{d}{dt} T_{n+1}(t) \tag{54}$$

The return response to a unit step of the mth integral is

$$x_m(t) = c_0 u_m(t) + c_1 u_{m-1}(t) + \ldots + c_{m-1} u_1(t) + T_{-m}(t) \tag{55}$$

where $T_{-m}(t)$ is the mth-integral transient term and c_m is the mth-integral initial-value coefficient. The symbol $u_1(t)$ represents a unit impulse at $t=0$, $u_2(t)$ represents a unit doublet at $t=0$, etc. The mth-integral transient term $T_{-m}(t)$ is obtained by replacing n by $-m$ in the expressions for the nth-derivative transient term $T_n(t)$. The initial value coefficients may be calculated by replacing s by $1/\lambda$ in $G_{1u}(s)$ and employing the relation

$$c_m = \frac{1}{m!}\left[\frac{d^m}{d\lambda^m} G_{1u}(1/\lambda)\right]_{\lambda=0} \tag{56}$$

The error response to a step of a derivative can be obtained readily from the basic relationship

$$x_e(t) = x_i(t) - x_0(t) \tag{57}$$

Substituting equation 50 into equation 57 gives the following expression for the error response to a unit step of the nth derivative

Table IV. Expansions for Sin qθ/Sin θ

q	sin qθ/sin θ
0	0
1	1
2	2ζ
3	$4\zeta^2-1$
4	$4\zeta(2\zeta^2-1)$
5	$16\zeta^4-12\zeta^2+1$

$$e_s(t)=c_0\frac{t^n}{n!}+c_1'\frac{t^{n-1}}{(n-1)!}+\ldots+c_n'+T_n'(t) \tag{58}$$

where $T_n'(t)$ is the nth-derivative transient term for the error response and c_n' is the nth-derivative steady-state coefficient for the error response, commonly called error coefficient. These are equal to

$$T_n'(t)=-T_n(t) \tag{59}$$

$$c_n'=-c_n, \text{ for } n>0 \tag{60}$$

$$c_0'=1-c_0 \tag{61}$$

They may also be computed from the error-transfer function G_{te} as follows:

$$c_n'=\frac{1}{n!}\left[\frac{d^n}{ds^n}G_{te}\right]_{s=0} \tag{62}$$

$$T_n'(t)=\sum_{l=1}^{L}K_{ln}'\epsilon^{s_l t} \tag{63}$$

$$K_{ln}'=\left[\frac{(s-s_l)}{(s_l)^n}G_{te}\right]_{s=s_l} \tag{64}$$

The error response to a step of an integral is of little consequence, because such an input represents a high-frequency disturbance to the system which the system cannot follow, and hence, ideally, does not follow. For such an input the concept of "error" actually should refer to the return response, because any return response of the system to a disturbance actually corresponds to an error.

Generally, the simplest way to calculate the steady-state and initial-value coefficients is by a long division expansion of the return transfer function G_{te}, or the error transfer function G_{te}, as shown in reference 1. They can also be obtained from the initial values of the transient terms as follows:

$$c_n=-T_n(0)=-\sum_{l=1}^{L}K_{ln} \tag{65}$$

$$a_m=T_{-m}(0)=\sum_{l=1}^{L}K_{l(-m)}, \text{ for } m>0 \tag{66}$$

Computation of Transient Coefficients from Asymptotic Plot

The magnitudes of the coefficients of the transient terms may be estimated from asymptotic plots, enabling one to estimate the major characteristics of the transient terms.

Derivation

The magnitude of the return-response coefficient K_{rn} given in equation 53 can be expressed as

$$|K_{ln}|=\text{asm}\left|\frac{(s-s_l)}{s^{n+1}}G_{tu}\right|_{s=s_l}\times \text{dev}\left|\frac{(s-s_l)}{s^{n+1}}G_{tu}\right|_{s=s_l} \tag{67}$$

$$=|s_l|^{-n}\times\text{asm}\,|G_{tu}(s_l)|\times \text{dev}\,|(s-s_l)G_{tu}|_{s=s_l} \tag{68}$$

Since the error-response transient term $T_n'(t)$ is equal to $-T_n(t)$, the coefficients of the error-response transient terms are equal to

$$K_{ln}'=-K_{ln} \tag{69}$$

Hence the magnitude of the error-response and return-response transient coefficients are equal. The error-response transient coefficients may be expressed in terms of the error-transfer function by

$$|K_{ln}'|=|s_l|^{-n}\times\text{asm}|G_{te}(s_l)|\times \text{dev}\,|(s-s_l)G_{te}|_{s=s_l} \tag{70}$$

Coefficients for Step Response

To illustrate the application of equations 68 and 70, the step response coefficient for the pole at ω_3' shall be computed. Fig. 7 shows the magnitude asymptote of G_{te} at ω_3' to be $-15.5\,dg$, which corresponds to a ratio of 0.029. Since there are no other break frequencies of G_{te} close to ω_3', a reasonably good approximation to the transient coefficient for the pole at ω_3' can be obtained by neglecting the deviations. Since $n=0$ for the step response, equation 70 gives for the step response transient coefficient for the pole at ω_3'.

$$|K'_{(s')0}|\cong\text{asm}\,|G_{te}(\omega_3')| =-15.5\text{ dg}=0.029 \tag{71}$$

If the exact value for this coefficient is desired, the deviations must be included. With the index line of the deviation plot of Fig. 4 placed at ω_3', the following deviations are read for the individual break frequencies of G_{te}: -0.45 dg at ω_1, -1.3 dg at ω_2, $+1.4$ dg at ω_4', -0.2 dg at ω_5, -0.05 dg at ω_6, $+0.05$ dg at ω_7', and for the two poles at ω_3' (having a damping ratio of 0.678) a deviation of 0.25 dg per pole. This gives a total deviation of -0.05 decilogs which is negligible with respect to the asymptote value, and hence may be neglected.

The magnitude of this coefficient for the pole at ω_3' can also be computed from G_{tu} but the effect of the deviation must be considered. At ω_3' the deviation is quite large because G_{tu} has a zero at $-\omega_3$, which is very close to the pole at $-\omega_3'$. The deviation for the zero $-\omega_3$ is equal to, by equation 12

$$\text{dev}=\left|\frac{\Delta s}{s+}\right|=\frac{(\omega_3-\omega_3')}{\omega_3}=\frac{0.08-0.0776}{0.08} =0.030 \tag{72}$$

The magnitude asymptote of G_{tu} at ω_3' is unity. If the deviations due to the other break frequencies of G_{tu} are neglected, the coefficient for the pole at ω_3' is then approximately

$$|K'_{(s')0}|\cong\text{asm}\,|G_{tu}(\omega_3')|\times\frac{(\omega_3-\omega_3')}{\omega_3} =(1)(0.030)=0.030 \tag{73}$$

This is in good agreement with the value given in equation 71, which was calculated from G_{te}.

This discussion shows that for a closed-loop pole below ω_c, the magnitude asymptote of G_{te} gives a much better indication of the values of the transient coefficients than does the magnitude asymptote of G_{tu}, because the closed-loop pole is so near a corresponding zero of G_{tu} that the deviation due to that zero is very large. Similarly, for a closed-loop pole above ω_c the magnitude asymptote of G_{tu} gives a much better indication of the values of the transient coefficients than does that of G_{te}, because the closed-loop pole is quite close to a corresponding zero of G_{te}.

To illustrate this point further, assume that except for the zero of G_{tu} at ω_3 all the break frequencies of G_{te} and G_{tu} are so far from ω_3' that their deviations at ω_3' are negligible. Then the deviation for G_{te} at ω_3' is unity and the deviation for G_{tu} at ω_3' is $(\omega_3-\omega_3')/\omega_3$. Hence the magnitudes of the step response transient coefficients for the error and return responses are equal to approximately

$$|K_{(s')0}|\cong\text{asm}\,|G_{tu}(\omega_3')|\frac{(\omega_3-\omega_3')}{\omega_3} \tag{74}$$

$$|K'_{(s')0}|\cong\text{asm}\,|G_{te}(\omega_3')| \tag{75}$$

Since these coefficients must be equal in magnitude, the right-hand expressions can be equated, which gives

$$\frac{\text{asm}\,|G_{tu}(\omega_3')|}{\text{asm}\,|G_{te}(\omega_3')|}=\frac{\omega_3}{|\delta_3|} \tag{76}$$

where the shift $|\delta_3|$ has been substituted for $(\omega_3-\omega_3')$. This result can be generalized as follows: The ratio of the magnitude asymptotes of the two closed-loop functions G_{te} and G_{tu} at a close-loop pole s_x' is approximately

$$\text{ratio of asymptotes}=\frac{|s_{xx'+}|}{|\delta_x|} \tag{77}$$

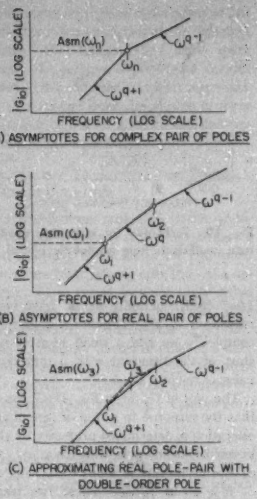


Fig. 14. Asymptotes for pair of poles close together

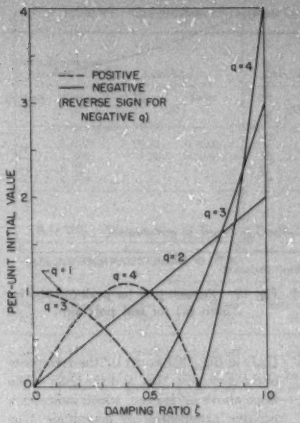


Fig. 15. Per-unit initial value for complex pole pair: $(-\sin q\theta/\sin q\theta)$

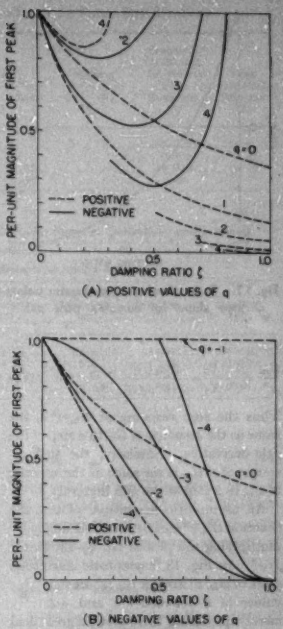


Fig. 16. Per-unit magnitude of first peak for complex pole pair

where δ_z is the shift of the closed-loop pole s_z' from the corresponding open-loop zero or pole s_z, and $|s_{zz}'_+|$ is the greater of the two frequencies $|s_z'|$ and $|s_z|$. This relation applies both for poles above and below ω_c: for a pole below ω_c the magnitude asymptote of G_{tu} is greater than that of G_{ti} by this ratio, and for a pole above ω_c the magnitude asymptote of G_{ti} is the greater by this ratio.

The approximation of equation 77 is illustrated graphically in Fig. 11 for the closed-loop poles at ω_2', ω_4', and ω_5'. This plot illustrates clearly how the same magnitudes of the transient coefficients can be calculated from both transfer functions G_{ti} and G_{tu} despite the fact that the magnitude plots are so different. It also explains why the magnitude asymptote of G_{tu} is more useful than that of G_{ti} for studying low-frequency effects; whereas the magnitude asymptote of G_{ti} is the more useful for studying high-frequency effects.

Whereas the magnitudes of transient coefficients can be computed from the values of the magnitude asymptotes, the signs of these coefficients can be determined from the slopes of the asymptotes. To show how this can be done, consider the magnitude asymptote plot of G_{tu} in Fig. 12(A) which has a single-order real pole at ω_1. If there are no other break frequencies of G_{tu} near the frequency ω_1, then G_{tu} can be approximated as follows in the frequency region near ω_1

$$G_{tu} = \frac{s^{q+1}\ \text{asm}\ |G_{tu}(\omega_1)|}{(s+\omega_1)\omega_1^q} \tag{78}$$

Making this approximation is equivalent to neglecting the deviations of the other break frequencies of G_{tu} at ω_1. The return step response transient coefficient for the pole at ω_1 is equal to

$$K_{10} = \left[\frac{(s+\omega_1)G_{tu}}{s}\right]_{s=-\omega_1} \tag{79}$$

Assuming the transfer function for G_{tu} in equation 78 gives

$$K_{10} = (-1)^q\ \text{asm}\ |G_{tu}(\omega_1)| \tag{80}$$

Thus if the slope parameter q (which is defined in terms of the asymptote slope above the pole) is even, the transient coefficient is positive, and if the slope parameter q is odd, the coefficient is negative.

Of course, equation 80 applies only for poles above ω_c, because for poles below ω_c the deviation cannot be neglected and the approximation for G_{tu} in equation 78 cannot apply. For poles below ω_c, one can apply the equivalent expression

$$K_{10}' = (-1)^{q'}\ \text{asm}\ |G_{ti}(\omega_1)| \tag{81}$$

where the slope parameter q' is again defined in terms of the asymptote above the frequency ω_1, as is shown in Fig. 12(B). It is important to note in employing the two transfer functions G_{ti} and G_{tu} that an error-response transient-coefficient is the negative of the corresponding return-response coefficient. Thus, the sign of the return-response transient-coefficient for a pole below ω_c is determined from the equation

$$K_{10} = -K_{10}' = -(-1)^{q'}\ \text{asm}\ |G_{ti}(\omega_1)| \tag{82}$$

These relationships between the magnitude asymptotes and the transient coefficients, of course, apply only for real single-order poles. Nevertheless, as is shown later, the magnitude asymptote also gives a good indication of the form of the transient component both for a pair of complex poles and for a multiple-order pole.

Computing Transient Coefficients for Steps of Derivatives and Integrals

To compute the transient coefficients for a step of the nth derivative, the same techniques may be used as for the step response, by employing a plot of the function G_{tu}/s^n (or G_{ti}/s^n) in place of the plot of G_{tu} (or G_{ti}). This is because G_{tu}/s^n is equal to the ratio of the transforms of the return signal and the nth derivative of the input

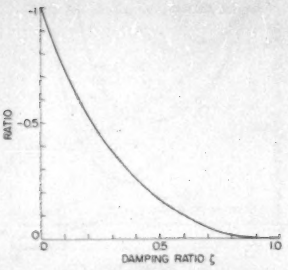


Fig. 17. Ratio of values at successive points of zero slope for complex pole pair

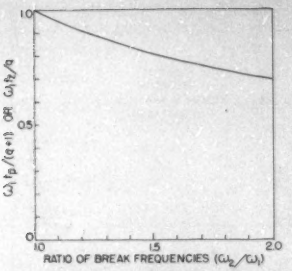


Fig. 18. Per-unit time for peak (t_p) and for zero (t_z) for real pole pair

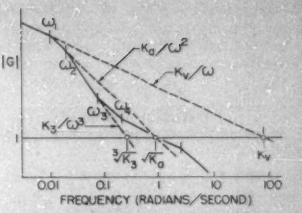


Fig. 19. Technique for estimating steady-state coefficients from asymptotic plot of G

$$\frac{G_{iu}}{s^n} = \frac{1}{s^n}\frac{X_u}{X_i} = \frac{\mathcal{L}[x_u(t)]}{\mathcal{L}[D^n x_i(t)]} \qquad (83)$$

Thus the step response of G_{iu}/s^n is the same as the response of G_{iu} to a step of the nth derivative. Similarly, the step response of $s^m G_{iu}$ is the same as the response of G_{iu} to a step of the mth integral.

An asymptotic magnitude plot of the function G_{iu}/s^n can be obtained simply by multiplying the plot of G_{iu} by the factor $1/\omega^n$. In Fig. 13, asymptotic magnitude plots of G_{iu}/s^n and/or G_{iu}/s^n are shown for values of n between -1 and $+3$. By merely inspecting these plots, a great deal of information can be determined concerning the shapes of the transient responses to an impulse, a step, a ramp, a step of acceleration, and a step of the third derivative.

Consider first the step response, which can be estimated from the plots of G_{iu} and G_{iu} in Fig. 13(A). Since the asymptote of G_{iu} at ω_3' is 1.4, the transient component for these poles has a maximum value of roughly 1.4. At ω_4', the asymptote of G_{iu} is equal to 0.4, indicating that the pole adds an exponential tail to the response with a magnitude of roughly 0.4 and a time constant of $1/\omega_4'$ or 3.5 seconds. At ω_3' the asymptote of G_{iu} is equal to 0.03, and hence this pole should add a very small but long exponential tail with a magnitude of roughly 0.03 and a time constant of $1/\omega_3'$ or 13 seconds. Since the order of the slope of G_{iu} above ω_4' is odd (one), while that above ω_3' is even (two), the pole at ω_4' adds a negative exponential to the error response while that at ω_3' adds a positive exponential. At ω_6' the asymptote of G_{iu} is 0.03, which indicates that the pole has little effect upon the step response, because the magnitude of the transient is so low and the time constant is quite short.

The shape of the ramp response can be estimated from the plots of G_{iu}/s and G_{iu}/s in Fig. 13(B). At ω_4' the asymptote of G_{iu} is roughly $1/\omega_c$, and the order of the slope above ω_4' is even, which indicates a positive exponential of magnitude $1/\omega_c$ and a time constant of $1/\omega_4'$. At ω_3' the asymptote is also roughly $1/\omega_c$ which indicates that the maximum value of the transient for the complex poles is also roughly unity. Actually the initial value of this component is negative, so that the transients for the poles at ω_3' and ω_4' subtract from one another to produce a net transient starting at about zero, rising to roughly $1/\omega_c$ and decaying back to zero. At ω_3' the asymptote is about $0.4/\omega_c$ and the order of the slope above ω_3' is odd, which indicates that the pole adds a long negative tail to the ramp error response with a magnitude of roughly $0.4/\omega_c$. The asymptote of G_{iu} at ω_6' is very small, indicating that that pole has negligible effect on the ramp response and on the responses to steps of higher derivatives.

Fig. 13(C) gives plots of G_{iu}/s^2 and G_{iu}/s^2, from which the responses to steps of acceleration and rate of acceleration can be estimated. Since these responses are determined primarily by the low-frequency poles, plots of G_{iu}/s^2 and G_{iu}/s^3 are not included.

For the response to a step of acceleration, consider the plot of G_{iu}/s^2. At ω_3' the asymptote is about $5/\omega_c^2$ and the order of the slope above ω_3' is even (zero), which indicates approximately a positive transient for the error response of magnitude $5/\omega_c^2$ and a 13-second time constant. At ω_4' the asymptote is also about $5/\omega_c^2$ and the order of the slope above ω_4' is odd (-1), thus indicating a negative exponential for the error response of magnitude $5/\omega_c^2$, and a time constant of 3.5 seconds. These two components thus subtract from one another, producing a net transient starting at about zero, rising to about $5/\omega_c^2$, and decaying back to zero. At ω_6' the asymptote is roughly $1/\omega_c^2$ (which actually should have been read from G_{iu}/s^2). This indicates that the complex poles add a small term at the start of the transient with a maximum value of roughly $1/\omega_c^2$.

The plot of G_{iu}/s^3 in Fig. 13(C) shows that the response to a step of rate of acceleration is determined primarily by the lowest-frequency pole, which is at ω_3'. At ω_3' the asymptote is about $60/\omega_c^3$ and at ω_4' it is about $15/\omega_c^3$. These two transients again subtract from one another to produce a net negative transient for the error response starting at about $-45/\omega_c^3$ and decaying back to zero roughly as an exponential with a time constant of 13 seconds.

For steps of derivatives higher than the third, it is obvious that the pole at ω_3' almost entirely determines the response.

The impulse response can be estimated from the plots of sG_{iu} and sG_{iu} in Fig. 13(D). The complex poles at ω_3' primarily determine the response. The asymptote of sG_{iu} at ω_3' is 1.8 ω_c, which indicates a maximum value of roughly 1.8 ω_c. The asymptote is 0.3 ω_c at ω_4', and the slope above ω_4' is odd (-3), which indicates a negative exponential component of roughly the magnitude 0.3 ω_c and a time constant of $1/\omega_4'$ or 0.12 second. Since the asymptote of G_{iu} and ω_3' and ω_4' is quite small, the poles at ω_3' and ω_4' have little effect on the impulse response and even less effect on the responses to steps of higher integrals.

These approximations, of course, are quite rough; nevertheless they give a

Table V. Comparison of Exact and Approximate Error Coefficients

Coefficient	Derivative	Values: Approximate	Values: Exact
c_0'		0	0
c_1'	velocity	0.0125 sec.	0.0125 sec
c_2'	acceleration	1.25 sec^2	1.063 sec^2
c_3'	rate of acceleration	62.5 sec^3	33.6 sec^3

good semiquantitative description of the important characteristics of the transient terms. Particularly, they show quite clearly the relative effect of the various poles on the different transient responses. To give a more exact description of these transient responses, a more detailed approximation is presented later.

Estimating Transients for Poles Close Together

The technique presented in the previous section is quite effective for estimating the transient components for real poles which are relatively distant from one another. However, it does not apply either to complex poles or to real poles quite close together. For such cases it is necessary to consider the set of poles as a unit, and to estimate the combined transient for the set. Consider first a method for estimating the transient for a pair of complex poles.

Complex Poles

The transient component for a pair of complex poles can be plotted quite accurately by finding the initial value, the values at points of zero slope, and the points where the transient is zero. These values can be estimated from the magnitude asymptotes provided the damping ratio of the poles is known.

Fig. 14(A) shows a portion of a magnitude asymptote plot of a function G_{to} having a pair of complex poles with the natural frequency ω_n and damping ratio ζ. For generality G_{to} is considered as the transfer function between given input x_i and a given output x_o, as was shown in Fig. 1(A), and may or may not represent the closed-loop transfer function of a feedback control loop. If the other poles and zeros of G_{to} are quite distant from the complex poles, so that their deviations may be neglected, G_{to} may be approximated as follows in the region near the complex poles

$$G_{to}=\frac{s^{q+1}\,\mathrm{asm}(\omega_n)}{(s^2+2\zeta\omega_n s+\omega_n{}^2)\omega_n{}^{q-1}} \tag{84}$$

For simplicity, $\mathrm{asm}(\omega_n)$ is used to designate asm $|G_{to}(\omega_n)|$, the magnitude asymptote of G_{to} at the frequency ω_n. The parameter q describes the slope of the asymptote at the pole, and is defined in Fig 14(A).

Appendix I shows that the transient portion of the step response for the approximation of G_{to} given in equation 84 is

$$x_{\sigma}(t)=(-1)^q\left(\frac{\omega_n}{\omega_0}\right)\mathrm{asm}(\omega_n)\epsilon^{-\zeta\omega_n t}\times \sin\,(\omega_0 t-q\theta) \tag{85}$$

where θ and ω_0 are defined by

$$\theta=\cos^{-1}\zeta \tag{86}$$

$$\omega_0=\omega_n\sin\theta=\omega_n\sqrt{1-\zeta^2} \tag{87}$$

To find the initial value of the transient, set $t=0$ in equation 85. Since ω_0/ω_n is equal to $\sin\theta$, this gives

$$x_o(0)=(-1)^{q+1}\,\mathrm{asm}\,(\omega_n)\left(\frac{\sin q\theta}{\sin\theta}\right) \tag{88}$$

Table IV lists simple expressions for $(\sin q\theta/\sin\theta)$ which are derived in Appendix I. Take the negative of these for negative values of q. Equation 88 differs by the factor $(\sin q\theta/\sin\theta)$ from equation 80, which gave the initial value for a real pole. This factor is plotted in Fig. 15 and is called the per-unit initial value for the complex pole pair. The total initial value is

$$x_o(0)=[(-1)^q\,\mathrm{asm}(\omega_n)] \quad \text{[per-unit initial value]} \tag{89}$$

To obtain the values of the transient at points of zero slope, set the derivative of equation 85 equal to zero. This gives

$$\tan\,(\omega_0 t-q\theta)=\tan\,\theta \tag{90}$$

which corresponds to

$$(\omega_0 t-q\theta)=\theta+p\pi \tag{91}$$

$$\omega_0 t=\theta(1+q)+p\pi \tag{92}$$

where p represents an integer. Hence one point of zero slope must occur at a time t_p given by

$$t_p=\theta(1+q)/\omega_0 \tag{93}$$

and the other points must be separated from this point by intervals of time given by

$$\Delta t=\pi/\omega_0 \tag{94}$$

Thus, points of zero slope occur at $(t_p+\Delta t)$, $(t_p+2\Delta t)$, $(t_p+3\Delta t)$, etc. Substituting equation 91 into equation 85 gives for the values of the transient at points of zero slope

$$x_o=(-1)^{q+p}\,\mathrm{asm}\,(\omega_n)\epsilon^{-\zeta\omega_n t} \tag{95}$$

If the value for t_p given by equation 93 is less than Δt but not less than zero, the first point of zero slope occurs at t_p. Substituting equation 93 into equation 95 gives for the value of the transient at the first point of zero slope

$$x_o=(-1)^q\,\mathrm{asm}\,(\omega_n)\epsilon^{-(1+q)\theta/\tan\theta} \tag{96}$$

since $\omega_0/\zeta\omega_n$ is equal to $\tan\theta$. If t_p is less than zero or greater than Δt, one must add or subtract multiples of Δt to t_p to determine the time for the first point of zero slope.

Plots of the per-unit values of the transient at the first points of zero slope are given in Figs. 16(A) and 16(B) for values of q between -4 and $+4$. To obtain the complete values, multiply the per-unit values by $(-1)^q\mathrm{asm}(\omega_n)$. Note that the curves have discontinuities at a number of points. At these points the first peak occurs at its initial value. For $q=-1$, t_p in equation 93 is always zero, which indicates that the first peak occurs at $t=0$ for all values of ζ.

To obtain the values of the transient at later points of zero slope, it can readily be shown that the ratio of the values at two successive points of zero slope is

$$-\epsilon^{-\zeta\omega_n\Delta t}=-\epsilon^{-\pi/\tan\theta} \tag{97}$$

Thus the transient at successive points of zero slope must reverse in sign, as would

Table VI. Computation of Transient Coefficients for Return Step Response

Break Frequency of Pole		Asm|G_{to}|		Asm|G_{to}|		Slope q		Transient Coefficient for Return Step Response	Time Constant, Seconds
Symbol (1)	Value (2)	Dg	Ratio (3)	Dg	Ratio (4)	G_{to}	G_{to} (5)	(6)	(7)
ω_1'	0.0776	−15.51	0.0292			2		−0.0292	12.9
ω_2'	0.288	− 3.95	0.404			1		0.404	3.47
ω_3'	8.39			−14.7	0.0342		−3	−0.0342	0.12
ω_4'	1.306			1.5	1.40		−1	−1.40	

Table VII. Computation of Transient Coefficients for Steps of Derivatives and Integrals

Pole (1)	Break Frequency (2)	Transient Coefficients for Return Response: Step (3)	Ramp (4)	Step of Acceleration (5)	Step of Third Derivative (6)	Impulse (7)	Doublet (9)
$-\omega_1'$	0.0776	(−0.0292)	0.376	−4.85	62.5	(0.0021)	(-1.6×10^{-4})
$-\omega_2'$	0.299	0.404	−1.40	4.85	−16.6	−0.115	(0.033)
$-\omega_3'$	8.39	(−0.034)	(0.004)	(-4.9×10^{-4})	(5.9×10^{-5})	0.285	−2.38
$-\omega_4'$	1.306	−1.40	1.07	−0.82	(0.63)	1.83	−2.38

Table VIII. Computation of Points on Transients Due to Pole at ω_3'

t/τ	t_z Seconds	Ramp	Step of Acceleration	Step of Third Derivative
0	0	0.376	−4.85	62.5
1	12.9	0.138	−1.78	23.0
2	25.8	0.051	−0.65	8.5

be expected. The ratio of the values at successive points is plotted in Fig. 17 as a function of ζ, and as the plot shows, the ratio is negative.

It is also helpful to know the points where the transient is zero. Setting the expression for the transient in equation 85 equal to zero gives

$$\sin(\omega_n t - q\theta) = 0 \tag{98}$$

or

$$\omega_n t = q\theta + p\pi \tag{99}$$

Hence a zero occurs at a time t_z given by

$$t_z = q\theta/\omega_0 \tag{100}$$

and at all values of time separated from t_z by the time interval Δt, given previously.

This method for plotting the transient component for a pair of complex poles is illustrated in more detail later when the transient responses for the illustrative system are calculated.

In calculating the transient component for the complex poles, all the deviations were neglected in the transfer function $G_{i\omega}$. For the sake of completeness it is desirable to consider a means for employing these deviations. Since the poles are underdamped, the deviation for $G_{i\omega}$ at either of these poles, in general, must be complex. Assume that the deviation for $G_{i\omega}$ at the complex pole in the upperhalf-plane is represented by a magnitude D and phase angle γ. The resultant transient response given in equation 85 becomes as follows when this deviation is included

$$x_o(t) = D(-1)^q\left(\frac{\omega_n}{\omega_0}\right)\operatorname{asm}(\omega_n)\epsilon^{-\zeta\omega_n t}\times \sin(\omega_0 + \gamma - n\theta) \tag{101}$$

Now define a deviation time-interval t_d by

$$t_d = \gamma/\omega_0 \tag{102}$$

Hence

$$\omega_0 t + \gamma = \omega_0(t + t_d) \tag{103}$$

Thus the deviation has the effect of causing a shift of the time axis by t_d. Including this deviation time in the exponential as well as in the sine expression gives for the transient

$$x_o(t) = [D\epsilon^{\zeta\omega_n t_d}](-1)^q\left(\frac{\omega_n}{\omega_0}\right)\times \operatorname{asm}(\omega_n)\epsilon^{-\zeta\omega_n t'}\sin(\omega_0 t' - n\theta) \tag{104}$$

where t' represents $t + t_d$.

Thus, the effect of the deviation may be included in plotting the transient by shifting the zero time axis to the right by the interval t_d, which reduces the size of the transient, and then building the size of the transient back up again by increasing its amplitude by the factor.

$$D\epsilon^{\zeta\omega_n t_d} = D\epsilon^{\gamma/\tan\theta} \tag{105}$$

Real Poles Close Together

If two real poles of a transfer function are moved closer and closer together, for coefficients for these poles become infinite, because the denominators of the coefficients contain as a factor the distance between the poles. In spite of this, however, the resultant transient does not become infinite because the individual transient coefficients must subtract from one another. Consequently, for two real poles that are close together it is best to consider the resultant transient of the two as a unit, just as for a pair of complex poles.

Fig. 14(B) shows a portion of the magnitude asymptote of a function $G_{i\omega}$ which has two real poles quite close together, with break frequencies at ω_1 and ω_2. The plot shows that $G_{i\omega}$ can be approximated as follows in the region near ω_1 and ω_2, provided there are no other poles and zeros of $G_{i\omega}$ close to that region.

$$G_{i\omega} \cong \frac{s^{q+1}\omega_2 \operatorname{asm}(\omega_1)}{(s+\omega_1)(s+\omega_2)\omega_1^q} \tag{106}$$

Note that the slope parameter q is defined in the same manner as for the pair of complex poles which was shown in Fig. 14(A).

Appendix II shows that the transient part of the unit step response of $G_{i\omega}$ given in equation 106 is

Table IX. Computation of Points on Transients Due to Pole at ω_1'

t/τ	t_z Seconds	Step	Ramp	Step of Acceleration	Step of Third Derivative	Impulse
0	0	0.404	−1.40	4.85	−16.6	−0.115
1	3.47	0.147	−0.51	1.78	− 6.13	−0.042
2	6.94	0.053	−0.188	0.65	− 2.26	−0.016

Table X. Computation of Points on Transients Due to Pole at ω_2'

t/τ	t_z Seconds	Impulse	Doublet
0	0	0.285	−2.38
1	0.12	0.105	−0.88
2	0.24	0.039	−0.33

$$x_o(t) = \frac{(-1)^q\omega_2 \operatorname{asm}(\omega_1)}{(\omega_2 - \omega_1)}\times \left[\epsilon^{-\omega_1 t} - \left(\frac{\omega_2}{\omega_1}\right)^q \epsilon^{-\omega_2 t}\right] \tag{107}$$

As for the case of complex poles, this transient can also be sketched readily if one computes the initial value, the points of zero slope, and the points where the transient goes through zero.

Setting $t = 0$ in equation 107 gives the initial value of the transient as

$$x_o(0) = (-1)^{q+1}\operatorname{asm}(\omega_1)\frac{\omega_2(\omega_2^q - \omega_1^q)}{\omega_1^q(\omega_2 - \omega_1)} \tag{108}$$

Now, it can easily be shown that

$$(\omega_2^q - \omega_1^q) = (\omega_2 - \omega_1)(\omega_2^{q-1} + \omega_2^{q-2}\omega_1 + \ldots + \omega_2\omega_1^{q-2} + \omega_1^{q-1}) \tag{109}$$

Substituting this expansion into equation 108 gives for the initial value of the transient for negative q

$$x_o(0) = (-1)^q \operatorname{asm}(\omega_1)[1 + (\omega_1/\omega_2) + (\omega_1/\omega_2)^2 + \ldots + (\omega_1/\omega_2)^{|q|-1}] \tag{110}$$

for positive q

$$x_o(0) = (-1)^{q+1}\operatorname{asm}(\omega_2)[1 + (\omega_1/\omega_2) + (\omega_1/\omega_2)^2 + \ldots + (\omega_1/\omega_2)^{q-1}] \tag{111}$$

For q = zero, the initial value is zero. Note that it is the greater of the two asymptote values at ω_1 and ω_2, which appears in these two expressions. For negative q, the asymptote at ω_1 is greater than that at ω_2, but for positive q the asymptote at ω_2 is the greater.

To find the points of zero slope, set the derivative of equation 107 equal to zero. This gives

$$\epsilon^{-\omega_2 t} = \left(\frac{\omega_1}{\omega_2}\right)^{q+1}\epsilon^{-\omega_1 t} \tag{112}$$

Substituting this back into the expression for the transient in equation 107 gives an exponential curve along which the points of zero slope must lie

$$x_o(t) = (-1)^q \operatorname{asm}(\omega_1)\epsilon^{-\omega_1 t} \tag{113}$$

Solving equation 112 for the time of occurrence t_p of these points of zero slope gives

$$\omega_1 t_p = (q+1)\left[\frac{\log_\epsilon(\omega_2/\omega_1)}{(\omega_2/\omega_1) - 1}\right] \tag{114}$$

The expression within the brackets is

plotted in Fig. 18 for values of the frequency ratio ω_2/ω_1 from 1 to 2.

By employing the following infinite series

$$\log_e Y = 1 - \frac{1}{2}(Y-1) + \frac{1}{3}(Y-1)^2 - \ldots, \quad \text{for } 0 < Y < 2 \quad (115)$$

Equation 114 can be expressed conveniently as a series

$$\omega_1 t_p = (1+q)\left[1 - \frac{1}{2}\left(\frac{\Delta\omega}{\omega_1}\right) + \frac{1}{3}\left(\frac{\Delta\omega}{\omega_1}\right)^2 + \ldots\right], \quad \text{for } -\omega_1 < \Delta\omega < \omega_1 \quad (116)$$

where $\Delta\omega$ represents the interval between the poles $(\omega_2 - \omega_1)$.

To find the points where the transient is zero, set equation 107 equal to zero. This yields the expression

$$\omega_1 t_z = q\left[\frac{\log_e(\omega_2/\omega_1)}{(\omega_2/\omega_1) - 1}\right] \quad (117)$$

The expression within the brackets is the same as the expression plotted in Fig. 18, which was used to determine the occurrence of points of zero slope.

Examination of equations 114 and 117 for t_p and t_z shows the following. There can be only one zero and one point of zero slope on the transient. If $q = 0$, the initial value of the transient is zero, and for $q = -1$, the initial value of the transient is the point of zero slope. For positive values of q the transient has a zero, but for negative values of q it has no zeros because equation 117 yields a negative value for t_z. Likewise, for q greater than -1, the transient has a point of zero slope at positive values of time, but no points of zero slope for values of q less than -1. The primary effect of these conditions is that the transient looks much like a real exponential for negative q, but has an overshoot for positive q.

If the two poles are quite close together it is often convenient to approximate them by a single double-order pole with a break frequency at the geometric mean of the two single-order break frequencies. The asymptote value for this equivalent double-order pole can be found by extending the asymptotes above and below the real pole pair until they intersect. This construction is illustrated in Fig. 14(C).

For a double-order pole with the break frequency ω_3 the initial value for the transient can be obtained by setting $\omega_1 = \omega_2 = \omega_3$ in equations 110 and 111. This gives, at all values of the slope q

$$x_o(0) = q(-1)^{q+1} \operatorname{asm}(\omega_3) \quad (118)$$

To obtain the values of time for zeros and points of zero slope, make this same substitution in equations 114 and 117 giving

Table XI. Points on Transient Due to Poles at ω_1'

Line	Items	Step	Ramp	Step of Acceleration	Impulse	Doublet
1	Transient coefficients	−1.40	+1.07	−0.82	1.83	−2.38
2	Slope q	−1	−2	−3	0	1
3	Per-unit initial value	1	1.34	0.80	0	−1
4	Initial value	−1.40	1.43	−0.656	0	2.38
5	Per-unit first peak	1	−0.103	−0.256	0.466	0.217
6	Value at first peak	−1.40	−0.11	0.21	0.855	−0.52
7	Value at second peak	+0.085	+0.0066	0.012	−0.051	0.031
8	t_p, seconds	0	−0.66	−1.72	0.86	1.72
9	t_z, seconds	−0.86	−1.72	−2.58	0	0.86
10	Time for first peak	0	2.42	1.56	0.86	1.72
11	Time for second peak	3.28	5.70	4.84	4.14	5.00
12	Time for first zero	2.42	1.56	0.70	0	0.86
13	Time for second zero	5.62	4.84	3.98	3.28	4.14

$$\omega_1 t_p = q + 1 \quad (119)$$

$$\omega_1 t_z = q \quad (120)$$

Substituting equations 119 into equations 113 gives for the value of the transient at the point of zero slope

$$x_o = (-1)^q \operatorname{asm}(\omega_1) e^{-(1+q)} \quad (121)$$

Estimating Values of Steady-state Coefficients

If an algebraic expression for the loop transfer function is known, it is fairly simple to compute the steady-state coefficients by a long-division expansion, as has been described. On the other hand, it is also desirable to be able to estimate these coefficients directly from an asymptotic magnitude plot of the loop transfer function in order to see more directly how to design a system to produce the desired values for these coefficients.

One means of computing the steady-state coefficients is to employ equation 65

$$c_n = -T_n(0) \quad (122)$$

which states that a steady-state coefficient is equal to the negative of the initial value of the transient term. If the transient term is primarily determined by one pole, this means of computing the steady-state coefficients is quite good; but if the transient has two large components of nearly equal magnitude which subtract from one another, it may be difficult to get a sufficiently exact value for the initial value of the transient term to compute the steady-state coefficient adequately. Hence it is desirable to consider other means for approximating the values of the steady-state coefficients.

Equation 62 expresses the steady-state error coefficients in terms of the error transfer function as

$$c_n' = \frac{1}{n!}\left[\frac{d^n}{ds^n} G_{is}\right]_{s=0} \quad (123)$$

If the loop has a single integration at low frequencies, as occurs in a great many feedback control systems, the loop transfer function can be expressed as

$$G = \frac{K_v[1-(s/s_{z1})][1-(s/s_{z2})]\ldots}{s[1-(s/s_{p1})][1-(s/s_{p2})]\ldots} \quad (124)$$

where K_v is the velocity constant, s_z represents a zero of G and s_p a pole of G. The error transfer function then may be expressed as

$$G_{is} = \frac{s[1-(s/s_{z1})][1-(s/s_{z2})]\ldots}{K_v[1-(s/s_{z1})][1-(s/s_{z2})]\ldots} \quad (125)$$

where s_z represents a closed-loop pole, Appendix IV shows that by applying equation 123 to equation 125, the first three derivative error coefficients are equal to

$$c_1' = \frac{1}{K_v} \quad (126)$$

$$c_2' = \frac{1}{K_v}\left[\sum\frac{1}{(-s_p)} - \sum\frac{1}{(-s_z)}\right] \quad (127)$$

$$c_3' = \frac{1}{2K_v}\left\{\left[\sum\frac{1}{(-s_p)} - \sum\frac{1}{(-s_z)}\right]^2 - \sum\frac{1}{(-s_p)^2} + \sum\frac{1}{(-s_z)^2}\right\} \quad (128)$$

Now, for the illustrative example

$$K_v = [sG]_{s=0} = \left[\frac{\omega_c\omega_2\omega_4}{\omega_1\omega_3}\right]_{s=0} \quad (129)$$

$$\sum\frac{1}{(-s_p)} = \frac{1}{\omega_1} + \frac{1}{\omega_2} + \frac{1}{\omega_3} + \frac{1}{\omega_3} \quad (130)$$

$$\sum\frac{1}{(-s_z)} = \frac{1}{\omega_2'} + \frac{1}{\omega_4'} + \frac{1}{\omega_c'} + \frac{1}{\omega_3'} + \frac{1}{\omega_3'}, \quad (131)$$

The closed-loop poles are designated by priming the corresponding frequency from which it is shifted. Thus $1/\omega_4'$ refers to the pole shifted from ω_4, which is combined with the pole shifted from ω_3 to form an underdamped pair of poles.

Since the reciprocals of the poles and zeros occur in the expressions for the error coefficients it is only the lowest-frequency poles and zeros which are of importance. Quite accurate approximations can be obtained by considering

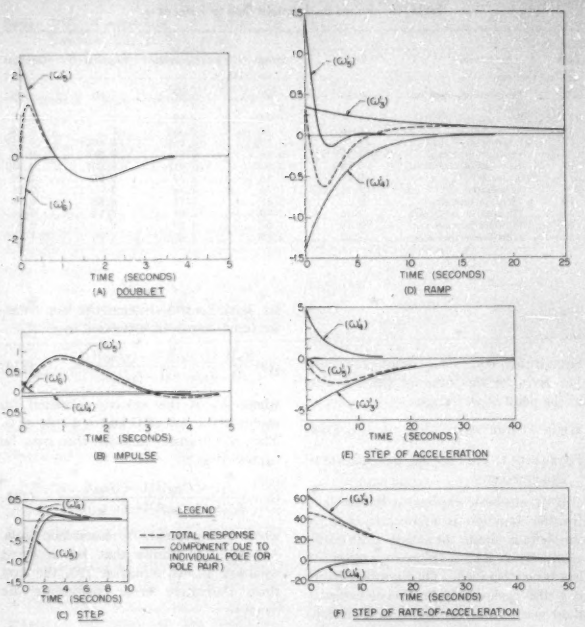


Fig. 20. Components of transient responses

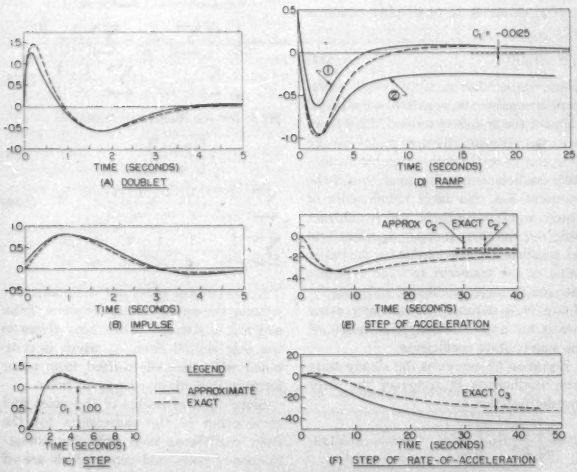


Fig. 21. Comparison of approximate and exact transient responses

only the lowest one in computing c_2' and only the two lowest in computing c_3'. This gives

$$c_1' \cong \frac{1}{K_v} = \frac{\omega_1\omega_2}{\omega_c\omega_3\omega_4} \tag{132}$$

$$c_2' \cong \frac{1}{K_v}\left[\frac{1}{\omega_2}\right] = \frac{\omega_1}{\omega_c\omega_3\omega_4} \tag{133}$$

$$c_3' \cong \frac{1}{2K_v}\left\{\left[\frac{1}{\omega_1}+\frac{1}{\omega_2}\right]^3 - \left(\frac{1}{\omega_1}\right)^3 - \left(\frac{1}{\omega_2}\right)^3\right\} \tag{134}$$

$$\cong \frac{1}{K_v\omega_3\omega_4} = \frac{1}{\omega_c\omega_3\omega_4} \tag{135}$$

These approximations can also be arrived at through intuitive reasoning by inspecting an asymptotic magnitude plot of the loop transfer function G, as is shown in Fig. 19. At very low frequencies G approaches the value K_v/s, where K_v is the velocity constant. Hence the low-frequency asymptote of G is a plot of K_v/ω. The frequency where the extension of this asymptote crosses the unity gain axis is the velocity constant K_v, as shown in the figure, because

$$\omega = K_v \text{ where } K_v/\omega = 1 \tag{136}$$

The value for K_v obtained from this construction is equal to the value that was given in equation 132.

Now, consider the result if the velocity constant were increased to infinity by decreasing the break frequency ω_1 to zero. At low frequencies G would approach the value K_a/s^2, where K_a would be the acceleration constant, and the asymptote at low frequencies would be an extension of the present asymptote between ω_1 and ω_2. Hence the frequency where the extension of this asymptote intersects the unity-gain axis would be equal to $\sqrt{K_a}$ because

$$\omega = \sqrt{K_a} \text{ where } K_a/\omega^2 = 1 \tag{137}$$

For the actual system, this point of intersection does not actually determine the acceleration constant. However, the value for K_a calculated from it is equal to the reciprocal of the approximation to the acceleration coefficient that was given above in equation 133. As shown in Fig. 19, the intersection occurs at 0.895 radian per second, which corresponds to an acceleration error coefficient of

$$c_3' = 1/(0.895)^2 = 1.25 \text{ seconds}^2 \tag{138}$$

To continue the line of reasoning, assume that both ω_1 and ω_2 are reduced to zero to produce an infinite acceleration constant. The loop transfer function G would approach the value K_3/s^3 at low frequencies, where K_3 would represent the rate-of-acceleration constant. The

Table XII. Exact Coefficients for Transient Responses

	Doublet	Impulse	Step	Ramp	Step of Acceleration	Step of Third Derivative
n	−2	−1	0	1	2	3
K_1	-1.63×10^{-4}	0.00210	−0.0271	0.349	−4.50	58.0
K_2	0.0484	−0.168	0.584	−2.02	7.03	−24.4
K_3	−2.96	0.353	−0.0421	0.00502	-5.98×10^{-4}	7.14×10^{-3}
K_3	−3.72	2.85	−2.18	1.67	−1.28	0.981
ϕ	−50.9°	−3.6°	43.7°	91.0°	138.3°	185.6°
c	0	0	1.000	−0.0125	−1.663	−33.6

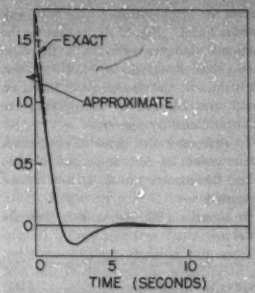


Fig. 22. Approximate and exact plots of ramp-response transient component for complex pole pair

low-frequency asymptote would be a plot of K_3/ω^3 and would be an extension of the present asymptote between ω_2 and ω_3. Hence, where the extension of this asymptote crosses, the unity gain axis is equal to $\sqrt[3]{K_3}$, as shown in the figure, because

$$\omega = \sqrt[3]{K_3} \text{ when } K_3/\omega^3 = 1 \qquad (139)$$

For the actual system the value of K_3 thus calculated from this intersection is not the actual rate-of-acceleration constant. Nevertheless it is equal to the reciprocal of the approximation to the rate-of-acceleration (or third-derivative) error coefficient that was given in equation 135. The intersection occurs at 0.252 radians per second which yields for the error coefficient

$$c_3' = 1/(0.252)^3 = 62.5 \text{ seconds}^3 \qquad (140)$$

The approximate error coefficients for the illustrative system computed by these techniques are listed in Table V along with the exact values.

Plotting Transients for Illustrative System

In Table VI the approximate values of the step response transient coefficients are computed from the values of the asymptotes of the closed-loop functions. Column 1 lists the symbols for the break frequencies for the closed-loop poles, and column 2 lists the numerical values. In column 3 are listed the values of the magnitude asymptote of G_{tt} at poles below ω_c, and in column 4 are listed the values of the magnitude asymptote of G_{tu} at poles above ω_c. These asymptote values are expressed both in decilogs and as ratios. In column 5 the order q of the asymptote slope is given in terms of G_{tt} for poles below ω_c and in terms of G_{tu} for poles above ω_c. The parameter q is defined for both a single-order pole and for a complex pair of poles, so that the magnitude asymptote of the corresponding function is proportional to $\omega^{(q-1)}$ just below the pole. In column 6 the approximate values are given for the transient coefficients of the return step response. The magnitude of each coefficient is equal to the value of the magnitude asymptote of the appropriate function G_{tt} or G_{tu} at the pole, and the sign of each is determined by the value of q. For the poles above ω_c (at ω_1' and ω_2') the slope parameter q of G_{tu} is odd, which indicates a negative coefficient for the return step-response. For the pole at ω_3', the slope parameter q of G_{tt} is even, which indicates a positive coefficient for the error step-response and hence a negative coefficient for the return step-response. Similarly, at ω_4', the slope parameter q of G_{tt} is odd, indicating a positive coefficient for the return step response. In column 7 are given the time constants of the transient components for the real poles, which are equal to the reciprocals of the corresponding break frequencies.

For the complex poles at ω_3', the coefficient computed in Table VI does not represent a true transient coefficient as do the coefficients for the real poles. Nevertheless, it does serve as a basis for computing the initial and peak values of the transient component, so it is convenient to consider it as a transient coefficient.

In Table VII the return response transient coefficients are computed for steps of derivatives and for steps of integrals, from the step response coefficients. Column 1 lists the poles, column 2 the break frequencies, and column 3 the step response transient coefficients computed in Table VI. Each coefficient is divided repeatedly by the negative of the corresponding break frequency to produce the coefficients for steps of velocity, acceleration, and rate-of-acceleration, listed in column 4, 5, and 6. By repeatedly multiplying each step response transient coefficient by the negative of the corresponding break frequency, the transient coefficients for an impulse and a doublet are computed and listed in columns 7 and 8. Thus, to generalize, a transient coefficient $c_{\dot{x}n}$ for the nth derivative (or $-n$th integral) is computed from the step response transient coefficient c_{x0} by

$$c_{\dot{x}n} = \frac{c_{x0}}{(-\omega_x)^n} \qquad (141)$$

where ω_x is the break frequency for the corresponding pole.

Comparing the separate transient coefficients in each column shows that for each transient the coefficients for some poles are so small with respect to the others that they have negligible effect. These coefficients are placed in parentheses, which indicate that they are to be neglected when the transients are plotted.

To sketch the real-exponential transient components attributable to the three real poles, the values of the exponentials are computed in Tables VIII, IX, and X at values of time equal to zero, one time constant, and two time constants. Consider Table VIII, for example, in which are computed the coefficients for the pole at ω_3'. Since the time constant is $1/\omega_3'$, or 12.9 seconds, points on the transients are computed for values of time of 0, 12.9 seconds, and 2(12.9) or 25.8 seconds. The values at 0 seconds are the transient coefficients computed in Table VII. Multiplying an initial value by $1/e$ or 0.37 gives the value at one time constant, and multiplying it again by 0.37 gives the value at two time constants.

To plot the transient components for the complex poles at ω_3', the initial value, the points of zero slope, and the points of zero value are calculated in Table XI for each transient. Line 1 lists the transient coefficients obtained from Table VII and line 2 lists the return-response slope parameter q. Line 3 lists the per-unit initial values obtained from Fig. 15 for the given values of q and the damping ratio 0.678 of the complex poles. Multiplying together lines 3 and 1 gives the actual initial values in line 4. Line 5 lists the per-unit values for the first peak which were read from Fig. 16. Multiplying these by the transient coefficients in line

1 gives the actual values for the first peak listed in line 6. Fig. 17 shows that for the particular damping ratio 0.678, the ratio of two successive peaks is -0.06. Hence, the values at the second peak listed on line 7 are obtained by multiplying the values in line 6 by -0.06.

To determine the times of occurrence of the points of zero slope and of zero value, the parameters t_p, t_z, and Δt are computed.

By equations 93, 100, and 94 they are equal to

$$t_p = (\theta/\omega_0)(1+q) = 0.880(1+q) \text{ second} \quad (142)$$

$$t_z = (\theta/\omega_0)q = 0.880\ q \text{ second} \quad (143)$$

$$\Delta t = (\pi/\omega_0) = 3.28 \text{ seconds} \quad (144)$$

since

$$\theta = \cos^{-1}\zeta = 47.3° = 0.826 \quad (145)$$

$$\omega_0 = \omega_n \sin\theta = 1.306 \sin 47.3° = 0.959 \quad (146)$$

The values of t_p and t_z are listed in lines 8 and 9 for the particular values of slope q. The values of t_p and t_z that are positive represent respectively the values of time for the initial peak and initial zero, which are listed in lines 10 and 12. For negative values of t_p and t_z, the time interval Δt is added to give the values in lines 10 and 12. To obtain the time for the second peak, given in line 11, add Δt to the values in line 10. Similarly, add Δt to the values in line 12 to obtain the time for the second zero given in line 13.

From the points computed in Tables VIII through XI, the individual transient components are plotted in Fig. 20. The component for each particular pole is labeled with the break frequency for that pole. In each figure the individual components are summed together to form the total transient term shown by a dashed curve.

When the steady-state coefficients are added to the transient terms, the resultant terms start at zero and approach the values of the steady-state coefficients at infinity. In Fig. 21 the approximate transient terms obtained from Fig. 20 are added to the approximate values of the steady-state return-response coefficients to obtain the composite transient terms. These are compared with the exact terms, shown as dashed curves, which are given by

$$[T_n(t)+c_n] = c_n + K_{2n}e^{-0.0776t} + K_{4n}e^{-0.295t} + K_{5n}e^{-0.295t} + K_{6n}e^{-0.295t}\sin(0.959t+\phi_n) \quad (147)$$

The coefficients for equation 147 are listed in Table XII.

To obtain the approximate transient responses for steps of position, velocity, and acceleration, the approximate values of error coefficients listed in Table V are used to determine the return-response steady-state coefficients. For the steps of position and velocity, these approximate coefficients are equal to the exact values. For the rate of acceleration, on the other hand, the approximate error coefficient given in Table V was not used because the resultant transient is almost monotonic, so that a better approximation is obtained by assuming that the steady-state coefficient is equal to the negative of the initial value of the transient portion. Thus, the resultant approximate curve in Fig. 21(F) for a step of rate of acceleration was obtained by shifting the transient portion vertically until the initial value was at zero.

The approximate step, impulse, and doublet responses of Fig. 21 agree very well with the exact responses, and there is also fairly good agreement in the responses to steps of acceleration and rate of acceleration. However, the ramp response approximation is not quite so good because the transient essentially consists of two transient components of almost equal magnitude subtracting from one another. This leads to fairly large errors even though the individual transient components of the ramp response are fairly well approximated. For example, Fig. 22 compares the approximate and exact plots of the ramp-response transient component for the complex pair of poles, and shows very good agreement except for a 20 per cent discrepancy of the initial value. On the other hand, a great deal of information can still be obtained from the approximate ramp response. In Fig. 21(D), there are plotted two approximate curves: approximation 1, which is formed by shifting the approximate transient curve vertically until the final value is equal to the steady-state coefficients and approximation 2, which is formed by shifting the curve until the initial value is zero. Since the individual transient components making up the response are approximated fairly well, the approximation 1 agrees quite well with the exact curve at the tail of the transient, and approximation 2 agrees quite well at the beginning.

Since the approximations of the transient responses were estimated from only the asymptote values of the closed-loop frequency responses, significantly better approximations could be obtained by computing the deviations at the poles. However, it should be emphasized that the main purpose for presenting the detailed approximation procedure is to provide the engineer with insight, whereby he can make rough but reliable estimates of the major characteristics of the transient responses merely by inspecting the asymptote frequency response plots. Consequently, it would rarely be necessary for the engineer to consider the deviations quantitatively, although he should understand their effect so that he can estimate it in a qualitative manner.

Conclusions

Simplicity in Applying Procedure

The techniques of this paper have been developed in great detail, so that they may provide high accuracy if sufficient pains are taken. However, in applying the techniques, one generally would not try to achieve high accuracy because the complexities involved tend to defeat the advantage of simplicity achieved by the approximations. On the other hand, it is necessary that the approximations be capable of extension to achieve high accuracy, so that the engineer can adequately estimate the accuracy of the approximations, and so that he can appreciate the general affect of any factors he may be disregarding.

Once the general approximation procedure has been understood, however, the engineer will find that he is usually able to obtain adequate design information with a rather crude application of the techniques. The simple rules for approximating transient responses that are presented in reference 3 illustrate how easy it is to do this. If the engineer appreciates the basis for the approximations, he should be able to interpret the results of a crude approximation in such a way that the crudeness does not limit the adequacy of his over-all design.

Summary of Approximations

As a summary, the major approximations of this paper are listed as follows.

Calculation of Closed-Loop Poles

Item 1. The closed-loop poles of a feedback control loop are roughly equal to the zeros of the loop transfer function G below gain crossover, the poles of G above gain crossover, plus a real pole at gain crossover.

Item 2. The actual closed-loop poles are shifted from the approximate values given in item 1 into the frequency regions where the magnitude asymptote is proportional to an odd power of frequency (i.e., the slope is an odd multiple of 10 decilogs per decade).

Item 3. The per-unit shift of an actual closed-loop pole from the approximation given in item 1 is roughly equal to the

value of the asymptote of $|G|$ at the approximate pole for poles above gain crossover, and for poles below gain crossover it is roughly equal to the asymptote of $|1/G|$ at the approximate pole.

Forming Asymptotic Magnitude Plots of Closed-Loop Functions

Item 4. As a first approximation, the magnitude of the return transfer function G_{tu} is equal to unity below gain crossover and to that of G above gain crossover; whereas the magnitude of the error transfer function G_{tu} is equal to that of $1/G$ below gain crossover and to unity above gain crossover.

Item 5. Exact magnitude plots of G_{tu} and G_{tu} can be developed readily from the approximations of item 4 by taking into account the shift of the exact closed-loop poles from the approximate poles given in item 1.

Plotting Transient Responses

Item 6. The magnitude of an individual transient component of the step response is roughly equal to the value of the magnitude asymptote of G_{tu} at the corresponding pole, for a pole below gain crossover, and for a pole above gain crossover it is roughly equal to the value of the magnitude asymptote of G_{tu} at the pole.

Item 7. Since the magnitude of G_{tu} is roughly equal to that of $1/G$ below gain crossover, and the magnitude of G_{tu} is roughly equal to that of G above gain crossover, item 6 can be more roughly approximated as follows: there is a component in the step response for each zero of G below gain crossover with a magnitude equal to the magnitude asymptote of $1/G$ at that zero; there is a component for each pole of G above gain crossover with a magnitude equal to the magnitude asymptote of G at that pole; and there is an exponential component with a magnitude equal to unity and a time constant equal to the reciprocal of the gain-crossover frequency.

Item 8. For a single-order pole, the sign for the transient coefficient is determined by the order of the asymptote slope immediately above the pole. If that asymptote is proportional to an odd power of frequency, the coefficient for the corresponding response is negative. Thus, if the magnitude asymptote of G_{tu} (or G) immediately above a high-frequency exact (or approximate) pole is proportional to an odd power of frequency or, in other words, has a slope which is an odd multiple of 10 decilogs per decade, the transient coefficient for the return step response is negative. Similarly, if the magnitude asymptote of G_{tu} (or $1/G$) immediately above a low-frequency exact (or approximate) pole is proportional to an odd power of frequency, the transient coefficient for the error step response is negative, and hence that for the return step response is positive. (Note that a high-frequency pole refers to one above gain crossover, whereas a low-frequency pole refers to one below gain crossover.)

Item 9. In considering the response to a step of the nth derivative (or $-n$th integral) plots of $G_{tu}s^n$ and G_{tu}/s^n should be considered in place of the plots of G_{tu} and G_{tu}.

Item 10. For a real closed-loop pole, the transient is a simple exponential with a time constant equal to the reciprocal of the break frequency of the pole.

Item 11. For a complex pair of closed-loop poles or for a pair of real poles quite close together, the transient is not so simple as for a real pole, but the magnitude asymptote of the appropriate function still gives a good indication of the size of the transient, and the time constant equal to the reciprocal of the break frequency still gives a good indication of the duration of the transient, provided the poles are reasonably well damped.

Item 12. For a pair of real poles close together or a pair of complex poles which are reasonably well damped, the transient component can be approximated fairly well by assuming that the pair of poles form a double-order pole. Define the slope parameter q, so that the magnitude asymptote above the pole pair is proportional to ω^{q-1}, while that below is proportional to ω^{q+1}. If the break frequency of the approximate double-order pole is designated ω_z, then the initial value of the transient is

$$x_o(0) = -q(-1)^q \operatorname{asm}(\omega_z) \tag{148}$$

The point of zero slope occurs on the exponential

$$x_o(t) = (-1)^q \operatorname{asm}(\omega_z)_e{}^{-\omega_z t} \tag{149}$$

and occurs at the time t_p

$$\omega_z t_p = 1 + q \tag{150}$$

There is a point of zero value occurring at the time t_z

$$\omega_z t_z = q \tag{151}$$

Thus, for q equal to zero the initial value of the transient is zero; for positive values of q the transient has an overshoot, and for negative values of q it is a monotonically decaying transient.

Steady-State Error Coefficients

Item 13. Fig. 19 shows how the error coefficients can be estimated readily from the low-frequency asymptotes of G for those derivatives at which the step response has two major transient components subtracting from one another. For steps of higher derivatives, however, the transient is primarily determined by the lowest frequency pole (or, sometimes, set of poles) and is essentially a simple exponential. For those transients, the initial value can be approximated quite readily, and this is equal to the negative of the corresponding steady-state coefficient. Thus, where the approximations of Fig. 19 do not apply, the steady-state coefficients can be determined readily from the values of the transient coefficients.

Loop Transfer Functions

The following items concern the differences in effect of characteristics of low-frequency, mid-frequency, and high-frequency portions of the loop transfer functions.

Item 14. In the region near gain crossover, the transfer function of practical feedback control loops must essentially approximate the function ω_c/s in order to achieve adequate stability under reasonable operating margins. Thus, in general, the rise time of the step response and the maximum error of the ramp response is roughly equal to $1/\omega_c$ and the maximum value of the impulse response is roughly ω_c.

Item 15. Adding a high-frequency lag to the loop transfer function has two primary effects:

1. It reduces the stability of the loop by adding phase lag at gain crossover.
2. It increases the filter action of the loop. A high-frequency lag has essentially the same filter action as a low-pass filter of that transfer function placed in series with the loop input.

Item 16. To add a low-frequency integral network to the loop transfer function has the two following primary effects:

1. It reduces the phase margin at gain crossover, so as to decrease stability.
2. It reduces the low-frequency components of the error.

The reduction in low-frequency components of the error signal is essentially the same as that which would occur if the original error signal were passed through a high-pass filter, having a transfer function equal to the reciprocal of that of the integral network.

Necessity for Approximations

The need for simple and reliable approximations in the field of feedback control cannot be overemphasized. Very often an engineer is faced with the problem of designing a feedback control system for which he has very poor information concerning system characteristics and performance requirements, and at the same time there are a great many alternative designs to choose from. For such applications a long, complicated analysis procedure is of little value. What is needed is a basic quantitative understanding of the fundamental dynamic considerations involved, and such an understanding must be based upon approximations.

Another reason why better approximations are needed is to provide improved criteria for specifying the dynamic characteristics of feedback control systems. Since more and more feedback control components are being packaged and refined as separate units, often the primary task of a control systems designer is to choose and put together the appropriate components. One of the most difficult parts of this task is that of determining what the specifications should be for the individual units and then finding the particular components which meet these specifications. Despite the general awareness that exists in the field today of the importance of dynamic considerations, in a great many cases it is nearly impossible to interpret dynamic specifications in a reasonably quantitative manner.

Consider, for example, the problem of specifying bandwidth. Almost every packaged instrument servomechanism sold today provides a specification of bandwidth, but it is generally difficult to find out what the specifications, such as "natural frequency," are, and it is usually impossible to relate a specification quantitatively to the dynamic requirements of one's particular application, even after the specification has been properly interpreted. Generally, the bandwidth specification that is used defines the magnitude curve of the closed-loop (return) transfer function, in terms of either the resonant frequency or the frequency where the magnitude curve falls below unity, −3 decibels, or −6 decibels. However, to define properly the speed of response of the system, the bandwith should be specified either in terms of the gain crossover frequency, or the phase of the return transfer function, or the magnitude of the error transfer function. In particular, the author has found that, along with the gain crossover frequency, the frequency where the phase lag of the return transfer function is 60 degrees (about one radian) is also a very useful measure of bandwidth, because it is generally closely equal to the gain crossover frequency.

Although there is a widespread belief that specifying the magnitude curve of the return transfer function in turn defines a reasonable specification of its phase curve, this is not so. In particular, magnetic amplifiers and other sampling devices, which introduce phase lag without attenuation, tend to produce a resonant frequency which is much higher than the gain crossover frequency. The wide-bandwith magnitude characteristic which results does not make the system faster; it merely makes it more responsive to high-frequency noise. In terms of the step response, it may cause a sharp rise or drop at the start of the response, but the actual rise time is still essentially a function of the gain crossover frequency.

To provide adequate means of specifying the dynamic performance of feedback control systems, better approximations are required. By means of such approximations, the seller can determine what specifications should be made, and the buyer can determine how to interpret these specifications in terms of his own requirements.

Many people have assumed that the problem of specification is solved by approximating the return transfer function with a simple quadratic of the form

$$G_{1\omega} = \frac{\omega_n^2}{s^2+2\zeta\omega_n s+\omega_n^2} \tag{152}$$

However, this "second-order system" approximation is usually inadequate and is frequently misused. Such a transfer function cannot even account for an additional integration at low frequencies: the velocity constant of the second-order system must always be equal to its asymptote crossover frequency $\omega_n/2\zeta$. An even more serious difficulty is that the parameters of the quadratic are usually determined by comparing its frequency response magnitude plot with that of the system. When this is done, the points of 60-degree phase lag may be tremendously different for the two functions, so that the resultant quadradic approximation does not even give a reasonable indication of the system's bandwidth.

The material in this paper, along with that in references 1, 2, and 3, is designed primarily to help fill this need for better approximations. The papers have developed a means for relating the important frequency response and transient response characteristics directly to the response to an arbitrary input. By expressing the dynamic behavior in terms of the response to an arbitrary input, the relationships are sufficiently general to be readily applied to any application. Thus, these approximations provide a frame of reference for explaining the dynamic aspects of feedback control in a simple, yet general, manner.

Appendix I. Expressing Transient for Complex Poles in Terms of Asymptote

Assume that the system transfer function, having a pair of complex poles at ω_n, can be approximated as follows in the region near ω_n

$$G_{1\omega} = \frac{s^{q+1}\,\mathrm{asm}\,(\omega_n)}{(s^2+2\zeta\omega_n s+\omega_n^2)\omega_n^{q-1}} \tag{153}$$

The expression $\mathrm{asm}(\omega_n)$ represents the value of the magnitude asymptote of $G_{1\omega}$ at the frequency ω_n. For a unit step input, the output $X_o(s)$ is

$$X_o(s) = \frac{s^q\,\mathrm{asm}\,(\omega_n)}{(s^2+2\zeta\omega_n s+\omega_n^2)\omega_n^{q-1}} \tag{154}$$

The transient part of the output is

$$X_o(s) = \frac{K}{s+\zeta\omega_n-j\omega_0} + \frac{K^*}{s+\zeta\omega_n+j\omega_0} \tag{155}$$

where

$$K = \left[\frac{s^q\,\mathrm{asm}\,(\omega_n)}{\omega_n^{q-1}(s+\zeta\omega_n+j\omega_0)}\right]_{s=-\zeta\omega_n+j\omega_0=-\omega_n/-\theta} \tag{156}$$

and

$$\theta = \cos^{-1}\zeta \tag{157}$$

$$\omega_0 = \omega_n\sqrt{1-\zeta^2} \tag{158}$$

Solving equation 155 gives

$$K = \frac{\mathrm{asm}\,(\omega_n)}{\omega_n^{q-1}}\,\frac{(-\omega_n/-\theta)^q}{2j\omega_0} \tag{159}$$

$$K = \frac{(-1)^q\omega_n\,\mathrm{asm}\,(\omega_n)}{\omega_0}\,\frac{\epsilon^{-j\theta q}}{2j} \tag{160}$$

Substituting equation 160 into equation 155 gives

$$X_o(s) = \frac{(-1)^q\omega_n\,\mathrm{asm}\,(\omega_n)}{\omega_0}\times \left[\frac{(\epsilon^{-j\theta q})/2j}{s+\zeta\omega_n-j\omega_0} - \frac{(\epsilon^{j\theta q})/2j}{s+\zeta\omega_n+j\omega_0}\right] \tag{161}$$

Take the inverse transform

$$x_o(t) = \frac{(-1)^q\omega_n\,\mathrm{asm}\,(\omega_n)}{\omega_0}\times \left[\frac{\epsilon^{-\zeta\omega_n t}\epsilon^{j(\omega_0 t-q\theta)} - \epsilon^{-\zeta\omega_n t}\epsilon^{-j(\omega_0 t-q\theta)}}{2j}\right] \tag{162}$$

$$x_o(t) = (-1)^q\,\mathrm{asm}\,(\omega_n)\epsilon^{-\zeta\omega_n t}\left[\frac{\omega_n}{\omega_0}\right]\times \sin(\omega_0 t-q\theta) \tag{163}$$

Appendix II. Expressing Transient for Two Real Poles, Close Together, in Terms of Asymptote

Assume that the system transfer function, having real poles at ω_1 and ω_2, can be approximated as follows in the region near ω_1 and ω_2.

$$G_{ia} = \frac{s^{q+1}\omega_2 \operatorname{asm}(\omega_1)}{(s+\omega_1)(s+\omega_2)\omega_1^q} \tag{164}$$

where ω_2 is greater than ω_1, and $\operatorname{asm}(\omega_1)$ represents the value of the magnitude asymptote of G_{ia} at the frequency ω_1. For a unit step input to G_{ia} the transform of the transient part of the output $X_o(s)$ is

$$X_o(s) = \frac{K_1}{s+\omega_1} + \frac{K_2}{s+\omega_2} \tag{165}$$

where

$$K_1 = \left[\frac{s^q\omega_2 \operatorname{asm}(\omega_1)}{(s+\omega_2)\omega_1^q}\right]_{s=-\omega_1} = \frac{(-1)^q\omega_2 \operatorname{asm}(\omega_1)}{(-\omega_1+\omega_2)} \tag{166}$$

$$K_2 = \left[\frac{s^q\omega_2 \operatorname{asm}(\omega_1)}{(s+\omega_1)\omega_1^q}\right]_{s=-\omega_2} = \frac{(-1)^q\omega_2^{q+1} \operatorname{asm}(\omega_1)}{(-\omega_2+\omega_1)\omega_1^q} \tag{167}$$

Comparing equations 166 and 167 shows that

$$K_2 = -(\omega_2/\omega_1)^q K_1 \tag{168}$$

Take the inverse transform of equation 165

$$x_o(t) = K_1 e^{-\omega_1 t} + K_2 e^{-\omega_2 t} \tag{169}$$

Substitute equations 166 and 168 into equation 169

$$x_o(t) = \frac{(-1)^q\omega_2 \operatorname{asm}(\omega_1)}{(\omega_2-\omega_1)} \times [e^{-\omega_1 t} - (\omega_2/\omega_1)^q e^{-\omega_2 t}] \tag{170}$$

Appendix III. Expansion of Expression Sin $(q\theta)$/Sin(θ)

To expand $\sin(q\theta)$, express it as the imaginary part of an exponential

$$\sin(q\theta) = Im\{e^{iq\theta}\} = Im\{(e^{i\theta})^q\} = Im\{(\cos\theta + j\sin\theta)^q\} \tag{171}$$

Now, by the binomial theorem

$$(\cos\theta + j\sin\theta)^q = \cos^q\theta + \frac{q}{1!}\cos^{q-1}\theta\, j\sin\theta + \frac{q(q-1)}{2!}\cos^{q-2}\theta\,(j)^2\sin^2\theta + \dots \tag{172}$$

which consists of $(q+1)$ terms, if q is a positive integer. Hence by equation 171, $\sin(q\theta)$ is the imaginary part of equation 172, which is

$$\sin(q\theta) = q\cos^{q-1}\theta\sin\theta - \frac{q(q-1)(q-2)}{2!}\cos^{q-3}\theta\sin^3\theta + \dots \tag{173}$$

which consists of $(q+1)/2$ terms if q is odd and $q/2$ terms if q is even. Factoring out $(\sin\theta\cos^{q-1}\theta)$, and expressing the series in a compact form, gives

$$\frac{\sin(q\theta)}{\sin\theta} = \cos^{q-1}\theta \sum_{k=0}^{k=\left\{\frac{q-1}{2}\right\}} \times \frac{(-1)^k q! \tan^{2k}\theta}{(q-2k-1)!\,(2k+1)!} \tag{174}$$

where the symbol $[(q-1)/2]$ represents the integer part of $(q-1)/2$. If we define the damping ratio ζ in terms of θ by

$$\zeta = \cos\theta \tag{175}$$

Then

$$\tan\theta = \sqrt{1-\zeta^2}/\zeta \tag{176}$$

and the expansion of equation 174 becomes

$$\frac{\sin(q\theta)}{\sin\theta} = \zeta^{q-1} \sum_{k=0}^{k=\left\{\frac{q-1}{2}\right\}} \times \frac{(-1)^k q! (\zeta^{-2}-1)^k}{(q-2k-1)!\,(2k+1)!} \tag{177}$$

Appendix IV. Derivation of Expansion for Error Coefficients

If a feedback control loop has one integration at zero frequency, which occurs in a great many cases, the error transfer function can be expressed in the form

$$G_{ie} = \frac{s[1-(s/s_{b1})][1-(s/s_{b2})]\dots}{K_v[1-(s/s_{a1})][1-(s/s_{a2})]\dots} \tag{178}$$

where K_v is the velocity constant, s_{a1}, s_{a2}, etc., are the closed-loop poles, and s_{b1}, s_{b2}, etc., are the poles of the loop transfer function G. The first three derivative error coefficients can be expressed as

$$c_1' = \left.\frac{G_{ie}}{s}\right|_{s=0} = \frac{1}{K_v} \tag{179}$$

$$c_2' = \left.\frac{d}{ds}\left[\frac{G_{ie}}{s}\right]\right|_{s=0} \tag{180}$$

$$c_3' = \frac{1}{2!}\left.\frac{d^2}{ds^2}\left[\frac{G_{ie}}{s}\right]\right|_{s=0} \tag{181}$$

To perform the differentiation indicated in equations 180 and 181 it is best to make a change of variables. Consider the function A which is a ratio of two products of variables.

$$A = \frac{x_1x_2x_3\dots}{y_1y_2y_3\dots} \tag{182}$$

By standard differentiation techniques, it can readily be shown that

$$dA = A\left[\sum\frac{dx}{x} - \sum\frac{dy}{y}\right] \tag{183}$$

$$d^2A = A\left\{\left[\sum\frac{dx}{x} - \sum\frac{dy}{y}\right]^2 - \left[\sum\left(\frac{dx}{x}\right)^2 - \sum\left(\frac{dy}{y}\right)^2\right] + \sum\frac{d^2x}{x^2} - \sum\frac{d^2y}{y^2}\right\} \tag{184}$$

To apply equations 183 and 184, note that x represents a factor of G_{ie} of the form $[1-(s/s_b)]$ and y represents a factor of the form $[1-(s/s_d)]$. Equations 180 and 181 for the acceleration and rate-of-acceleration error coefficients become

$$c_2' = \frac{1}{K_v}\left[\sum\frac{1}{(-s_b)} - \sum\frac{1}{(-s_d)}\right] \tag{185}$$

$$c_3' = \frac{1}{2K_v}\left\{\left[\sum\frac{1}{(-s_b)} - \sum\frac{1}{(-s_d)}\right]^2 + \sum\frac{1}{(-s_b)^2} - \sum\frac{1}{(-s_d)^2}\right\} \tag{186}$$

References

1. A Simple Method for Calculating the Time Response of a System to an Arbitrary Input, O. A. Biernson. *AIEE Transactions*, vol. 74, pt. II, 1955, pp. 227–45.

2. Estimating Transient Response from Open-Loop Frequency Response, George A. Biernson. *Ibid.* (Jan. 1956 section), pp. 388–403.

3. Quick Methods for Evaluating the Closed-Loop Poles of Feedback Control Systems, George A. Biernson. *Ibid.*, vol. 72, pt. II, May 1953, pp. 53–70.

Envelope Transfer Function Analysis in A-C Servo Systems

M. PANZER
NONMEMBER AIEE

IN THE ANALYSIS of a-c servo systems for stability and performance, attention is centered upon the carrier "envelope," i.e., the signal proper. Extension of the convenient transfer function description used in direct-signal (low-pass) servo systems thus requires that the dynamic behavior of each element in the a-c system be described in terms of its effect on the carrier envelope by "envelope transfer functions" which are functions of signal frequency. In general, three types of elements are encountered in a-c servo systems: type-1 elements in which both input and output are modulated carriers; type-2 elements which have inputs at signal frequencies and outputs which are modulated carriers, and are called "modulators"; and type-3 elements which accept modulated carriers as inputs, produce outputs at signal frequencies, and are called "demodulators." Both type-2 and type-3 elements readily lend themselves to simple envelope transfer function description if their dynamic behavior occurs on the signal side. Type-1 elements, however, have known transfer functions in terms of actual frequency, rather than signal frequency, and the envelope or signal transfer function must be derived.

In the case of band-pass type-1 elements, such as carrier-tuned networks inserted for equalization purposes, the derivation of the envelope transfer function is commonly accomplished through the use of the "narrow-bandwidth" approximation. Some attention[1,2] has been given to the problem of describing the envelope transmission in general terms applicable as well to high- or low-pass type-1 elements, but the work has involved unwieldy methods or a considerable degree of approximation. More recently, Chang[3] has presented the method of the "complex data transfer function" which provides a solution to the problem, although Chang's paper is relatively little concerned with the theoretical background and the treatment of the problem of cascading.

It is the purpose of this paper to unify and generalize the envelope transfer function technique, extending it to type-1 elements of all sorts. By using the Laplace transform, theorems are derived for obtaining the exact envelope transfer functions in terms of signal frequency, from the exact envelope transfer functions in terms of actual frequency. These permit analysis of cascaded elements and closed loops employing elements of all three types in any configuration. Inasmuch as it is not possible to obtain the exact envelope transfer function of a chain of cascaded type-1 elements by multiplying individual envelope transfer functions together, an approximation technique is devised for doing this with high accuracy. Finally, the envelope transfer functions of common type-1 elements are obtained and discussed in this paper.

Paper 56-792, recommended by the AIEE Feedback Control Systems Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Summer and Pacific General Meeting, San Francisco, Calif., June 25-29, 1956. Manuscript submitted October 24, 1955; made available for printing May 8, 1956.

M. PANZER is with Polytechnic Institute of Brooklyn, Brooklyn, N. Y.

Basic Theory

When a suppressed modulated carrier with a sinusoidal envelope is passed through a type-1 element, the two side-band frequencies which comprise the modulated carrier will, in general, undergo different attenuation and phase shift. It is then easy to show, through the use of either phasors or trigonometric identities, that the output of the element may be written as the sum of two suppressed modulated carriers which are in quadrature, 90 degrees out of carrier phase. It is further demonstrable that either of the two component carriers may be chosen to have an arbitrary phase shift with respect to the input carrier, and that the envelopes of the carriers will be affected by the choice of the carrier phase shift.

Now, this result may be extended by superposition to the case where the input is a carrier with an envelope that is periodic and has a line spectrum, and further extended by the transition to continuous spectra to inputs which are carriers with nonperiodic envelopes. The fact that one more superposition extends the result to inputs consisting of two modulated carriers in quadrature, assures that the output of a type-1 element anywhere in an a-c servo system may be represented as two suppressed modulated carriers in quadrature, with either carrier phase selected as desired. This representation is convenient and lends itself especially well to describing cascading with type-3 elements, since these latter are "phase-sensitive" and respond maximally to a particular carrier phase and not at all to the respective quadrature. The method is developed below.

Laplace Transform Description of Suppressed Carrier Amplitude Modulation

Consider a suppressed modulated carrier defined by

$$e_{cm} = e_m e_c = e_m \cos(\omega_c t + \theta) \tag{1}$$

where e_m is the signal and e_c the carrier. The Laplace transform of e_{cm} is given by

$$\begin{aligned} E_{cm} &= \mathcal{L} e_{cm} = \mathcal{L} e_m \cos(\omega_c t + \theta) \\ &= \mathcal{L} e_m Re_t[\epsilon^{i(\omega_c t + \theta)}] \\ &= \int_0^\infty e_m Re_t[\epsilon^{i(\omega_c t + \theta)}] \epsilon^{-st} dt \\ &= Re_t[\epsilon^{i\theta} \int_0^\infty e_m \epsilon^{-(s - i\omega_c)t} dt] \\ &= Re_t \epsilon^{i\theta} E_m(s - i\omega_c) \end{aligned} \tag{2}$$

where

$$E_m(s) = \mathcal{L} e_m(t) \tag{3}$$

In the foregoing, s describes actual or side-band frequency; i is used rather than j to avoid errors in the application of the operator Re_t, since s itself is complex. It is, of course, possible to write the cosine function as a sum of conjugate exponentials, but the use of the operator Re_t is more convenient.

If E_{cm} is the input to a type-1 element whose transfer function is $G(s)$, the output transform E_{cm}' is

$$E_{cm}' = Re_t G(s) \epsilon^{i\theta} E_m(s - i\omega_c) \tag{4}$$

The application of the inverse Laplace transform gives the output time function e_{cm}' as

$$e_{cm}' = \frac{1}{2\pi j} \int_{\sigma + j\infty}^{\sigma - j\infty} Re_t[G(s) \epsilon^{i\theta} E_m(s - i\omega_c)] \epsilon^{st} ds \tag{5}$$

Now let

$$s - i\omega_c = p \tag{6}$$

where p describes signal frequency; then

$$\begin{aligned} e_{cm}' &= \frac{1}{2\pi j} \int_{\sigma - i\omega_c + j\infty}^{\sigma - i\omega_c - j\infty} Re_t \times \\ &\quad [G(p + i\omega_c) \epsilon^{i(\theta + \omega_c t)} E_m(p)] \epsilon^{pt} dp \\ &= Re_t \epsilon^{i(\theta + \omega_c t)} \frac{1}{2\pi j} \int_{\sigma - i\omega_c + j\infty}^{\sigma - i\omega_c - j\infty} \times \\ &\quad G(p + i\omega_c) E_m(p) \epsilon^{pt} dp \end{aligned} \tag{7}$$

At this point it is convenient to introduce Φ such that

$$e_{em}' = Re_t e^{j(\theta+\omega_c t+\Phi)} \frac{1}{2\pi j} \int_{\sigma-i\omega_c-j\infty}^{\sigma-i\omega_c+j\infty} \times \frac{G(p+i\omega_c)}{e^{i\Phi}} E_m(p) e^{pt} dp \quad (8)$$

where Φ is an arbitrary phase angle discussed in the following.

If equation 8 is expanded in terms of real and imaginary parts with respect to i, there results

$$e_{em}' = \cos(\omega_c t+\theta+\Phi) \times \left[\frac{1}{2\pi j} \int_{\sigma-i\omega_c-j\infty}^{\sigma-i\omega_c+j\infty} Re_t \times \left(\frac{G(p+i\omega_c)}{e^{i\Phi}}\right) E_m(p) e^{pt} dp\right] \quad (9)$$

$$-\sin(\omega_c t+\theta+\Phi) \times \left[\frac{1}{2\pi j} \int_{\sigma-i\omega_c-j\infty}^{\sigma-i\omega_c+j\infty} Im_t \times \left(\frac{G(p+i\omega_c)}{e^{i\Phi}}\right) E_m(p) e^{pt} dp\right]$$

Envelope Transfer Functions

The envelope transfer functions may be obtained directly from equation 9. Evidently e_{em}' consists of two suppressed modulated carriers in quadrature. The phase angle Φ is the shift between the input carrier phase and an arbitrarily selected carrier phase in the output. In the bracketed terms which represent the envelopes, the presence of ω_c in the limits of integration does not affect the value of the integral because of the infinite limits, so that the bracketed terms are seen to have precisely the form of the inverse Laplace transform. Thus the output envelope transforms must be $E_m(p)\ Re_t\ [G(p+i\omega_c)/e^{i\Phi}]$ and $E_m(p)\ Im_t\ [G(p+i\omega_c)/e^{i\Phi}]$ respectively; and since the input envelope transform is $E_m(p)$, it follows that the envelope transfer functions are

$$G_{mc}(p) = Re_t[e^{-i\Phi} G(p+i\omega_c)] \quad (10A)$$

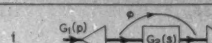
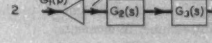
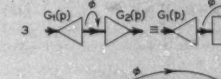
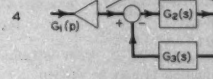
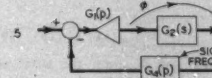
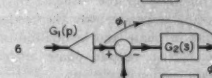
$$G_{qc}(p) = Im_t[e^{-i\Phi} G(p+i\omega_c)] \quad (10B)$$

Furthermore, if $\Phi=0$, and G_{mc} and G_{mq} are added to form a single complex function $\mathcal{G}$,

$$\mathcal{G} = G(p+i\omega_c) = G_{mc} + iG_{qc}, (\Phi=0) \quad (10C)$$

The transfer function $\mathcal{G}$ may be called the "entire" envelope transfer function which characterizes the type-1 element $G(s)$, since the real and imaginary parts of the complex product $E_m(p)\mathcal{G}$ represent respectively the transfer functions of the envelopes of both carriers in the output (for $\Phi=0$).

Table I. Over-all Envelope Transfer Functions

NC	CONFIGURATION	OVER-ALL ENVELOPE TRANSFER FUNCTION
1		$G_1(p)G_2(p)Re_t[e^{-i\Phi}\delta_2]$
2		$G_1(p)G_2(p)Re_t[e^{-i\Phi}\delta_2\delta_3]$
3		$G_1(p)G_2(p)Re_t[e^{-i\Phi}1] = G_1(p)G_2(p)\cos\Phi$
4		$G_1(p)G_4(p)Re_t\left[\frac{e^{-i\Phi}\delta_2}{1+\delta_2\delta_3}\right]$
5		$\frac{G_1(p)G_3(p)Re_t[e^{-i\Phi}\delta_2]}{1+G_1(p)G_3(p)G_4(p)Re_t[e^{-i\Phi}\delta_2]}$
6		$\frac{G_1(p)G_3(p)Re_t[e^{-i\Phi_1}\delta_2]}{1+G_3(p)G_4(p)Re_t[e^{-i\Phi_2}\delta_5\delta_2]}$
7		$\frac{G_1G_3G_4G_5G_7Re_t[e^{-i\Phi_1}\delta_2]Re_t[e^{-i\Phi_2}\delta_6]}{1+G_3G_4G_5G_7G_8Re_t[e^{-i\Phi_1}\delta_2]Re_t[e^{-i\Phi_2}\delta_6]}$

Envelope Transfer Functions of Cascaded Elements

Although there are several possible configurations of the three types of elements in cascade, they all may be readily described in terms of over-all envelope transfer functions once it is possible to describe the envelope effect of cascading type-1 elements. It may be readily shown that, with the approach just given, two type-1 elements $G_1(s)$ and $G_2(s)$, in cascade are characterized by an entire envelope transfer function which is the product of the individual entire envelope transfer functions

$$\mathcal{G}_{\text{cascaded}} = \mathcal{G}_1\mathcal{G}_2 \quad (11A)$$

It follows that the over-all envelope transfer function between the input and that carrier component of the output which is shifted by an angle Φ from the input carrier phase may be found from equation 10(A)

$$G_m(\text{cascaded}) = Re_t[e^{-i\Phi}\mathcal{G}_1\mathcal{G}_2] = Re_t[e^{-i\Phi}G_1(p+i\omega_c)G_2(p+i\omega_c)] \quad (11B)$$

Equations 10 and 11 enable over-all envelope transfer functions to be written down from inspection. To illustrate the procedure, Table I lists over-all envelope transfer functions for a few configurations. The symbols used for the different element types are shown in Fig. 1. The angles Φ in Fig. 1 and Table I are the phase shifts between an input carrier (at the tail end of the arrow) and the carrier component of proper phase for optimum response at the input to the type-3 element (at the head end of the arrow).

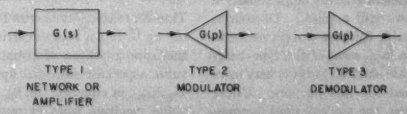


Fig. 1. Symbols for the several elements

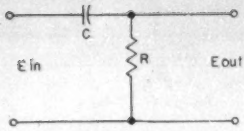


Fig. 2. C/R coupling circuit

Approximations for Cascaded Type-1 Elements

Although equation 11(A) provides an analytic means for describing the cascading of type-1 elements, the usefulness of the actual envelope transfer function found by taking the real part of the complex product of entire envelope transfer functions, as in equation 11(B), is limited in two ways. There is first the considerable amount of algebra involved in the application of equation 11(B), even to a few simple elements; but more important is the fact that simple superposition by the product of individual transfer functions is not possible. Thus it is impossible to single out a particular element of a cascaded chain, and in a simple way assess the effect of this element on the system performance; instead, it is necessary to view the cascaded structure as a whole. The feasibility of any simple evaluation of the contribution of a particular element to the over-all envelope transfer function of the cascade diminishes very quickly as the number of cascaded elements increases.

In view of this difficulty, it is fortunate that the envelope transfer function of cascaded type-1 elements may be reasonably well approximated as a product of individual element transfer functions, in the usual practical cases where type-1 elements are used. The approximation is made possible by the fact that the usual cascade of type-1 elements terminated by a type-3 element is so arranged that the carrier phase of the output of the last type-1 element for steady-state carrier conditions (zero signal frequency) is precisely the optimum value required for maximum response of the type-3 element. (In some systems, practical requirements may dictate that a small phase shift be tolerated. If the shift is small enough so that the system performance is not materially degenerated, the approximation which follows still applies.) Of course, this is desirable for best system performance and, in fact, some of the type-1 elements ("phase shift networks") may have been inserted for this purpose. Under this condition, the angle Φ used in equations 10 and 11(B) is simply the total carrier phase shift through the cascade of elements for steady-state carrier conditions, and thus may be written

$$\Phi = \Phi_c = \sum_n \Phi_{nc} \tag{12}$$

where Φ_{nc} is the carrier phase shift through the nth element under steady-state conditions. This permits equation 11(B) to be rewritten (extended to the case of any number of elements)

$$G_{m\,(\text{cascaded})} = Re_t \prod_n \mathcal{G}_n e^{-i\Phi_{nc}} = Re_t \prod_n \mathcal{G}_{nc}, \tag{13}$$

where

$$\mathcal{G}_{nc} = \mathcal{G}_n e^{-i\Phi_{nc}} \tag{14}$$

Using the value of $\mathcal{G}_n$ from equation 10(C), equation 14 may be written

$$\mathcal{G}_{nc} = e^{-i\Phi_{nc}} G_n(p + i\omega_c) = G_{nmc}(p) + iG_{nqc}(p) \tag{15}$$

where G_{nmc} and G_{nqc} are the envelope transfer functions of the element $G_n(s)$ corresponding respectively to the carrier components in phase and in quadrature with the carrier output phase under steady-state carrier conditions.

For $p=0$, it follows from the definition of Φ_{nc} that

$$G_n(0 + i\omega_c) = \left| G_n(i\omega_c) \right| e^{i\Phi_{nc}} \tag{16}$$

By substituting equation 16 in equation 15, it is clear that

$$G_{nmc}0 = \left| G_n(i\omega_c) \right| \tag{17C}$$

$$G_{nqc}0 = 0 \tag{17B}$$

Thus $G_{nqc}(p)$ has at least one zero at the origin. As $p = i\omega_m$ increases from zero, the amplitude of $G_{nqc}(p)$ initially increases no faster than linearly with ω_m. For sufficiently small ω_m, then, it is seen that

$$\left| G_{nmc}(i\omega_m) \right| >> \left| G_{nqc}(i\omega_m) \right|, \omega_m << \omega_c \tag{18A}$$

Of course, the largest value of ω_m which constitutes a sufficiently small value for which the approximation of equation 18(A) is still useful, is determined by the gain factors and the locations of the poles and zeros of both $G_{nmc}(p)$ and $G_{nqc}(p)$. Although no general discussion of this problem will be attempted here, for the representative networks analyzed later it is found that

$$\left| \frac{G_{qc}(i\omega_m)}{G_{mc}(i\omega_m)} \right| \leq \frac{\omega_m}{\omega_c}, \quad \frac{\omega_m}{\omega_c} \leq 0.15 \tag{18B}$$

Thus for reasonably narrow band systems, say up to 15 per cent of carrier frequency, the approximation of equation 18(A) is valid, especially since the upper bound of $|G_{qc}/G_{mc}|$ in equation 18(B) is a limiting value resulting from extreme parameter values which are not useful in practice.

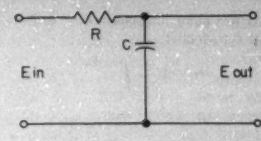


Fig. 3. R/C filter or phase shift circuit

Using the approximation of equation 18(A), equation 15 may be written

$$\mathcal{G}_{nc} \doteq G_{nmc}(p), \quad \frac{\omega_m}{\omega_c} \leq 0.15 \tag{19}$$

so that equation 13 becomes approximately

$$G_{m\,(\text{cascaded})} \doteq \prod_n G_{nmc}(p), \quad \omega_m/\omega_c \leq 0.15 \tag{20}$$

It is notable that the error in the approximation of equation 20 does not increase the number of cascaded elements is increased; in fact, the error decreases rapidly. This may be readily shown by considering the error of equation 20 in the case of two cascaded elements $G_1(s)$ and $G_2(s)$. For these two elements, the exact value of the cascaded G_m is, from equations 13 and 15,

$$G_m(\text{exact}) = Re_t[(G_{1mc} + iG_{1qc})(G_{2mc} + iG_{2qc})] = G_{1mc}G_{2mc} - G_{1qc}G_{2qc} \tag{21A}$$

whereas the approximate value is

$$G_m(\text{approx}) = G_{1mc}G_{2mc} \tag{21B}$$

The fractional error may be reasonably defined and found as

$$\text{Error} = \left| \frac{G_m(\text{exact}) - G_m(\text{approx})}{G_m(\text{approx})} \right| = \left| \frac{G_{1qc}G_{2qc}}{G_{1mc}G_{2mc}} \right| = \left| \frac{G_{1qc}}{G_{1mc}} \right| \left| \frac{G_{2qc}}{G_{2mc}} \right| \tag{22}$$

so that the over-all fractional error is the product of the individual fractional errors. If the approximation error is as much as 15% for each element, the over-all error is only about 2% for the two elements in cascade. Since additional cascading will reduce the error still further, the approximation of equation 20 may be used with equanimity.

If the quadrature envelope transfer function is desired, as well as the in-phase, it is easy to shown that G_q(cascaded) is given by

$$G_q(\text{cascaded}) = Im_t \prod_n \mathcal{G}_{nc} \tag{23}$$

It is unfortunately not possible to obtain an approximation for G_q (cascaded) as the product of individual element transfer functions. A kind of approximation is, of course, possible, as follows. Equation 23 may be written

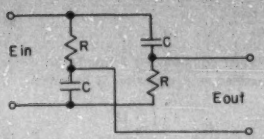


Fig. 4. Bridge phase shift circuit

$$G_g(\text{cascaded}) = Im_i \prod_n (G_{nmc} + iG_{nqc})$$
$$= Im_i \prod_n \left(G_{nmc}\left(1 + i\frac{G_{nqc}}{G_{nmc}}\right)\right)$$
$$= Im_i\left(\prod_n G_{nmc} \prod_n \left(1 + i\frac{G_{nqc}}{G_{nmc}}\right)\right) \quad (24)$$

Since $\prod_n G_{nmc}$ is itself the very good approximation for G_m(cascaded), and since

$$\prod_n \left(1 + i\frac{G_{nqc}}{G_{nmc}}\right) = 1 + i\sum_n \frac{G_{nqc}}{G_{nmc}} + \text{cross product terms} \quad (25)$$

where the cross product terms are of much smaller amplitude and may be neglected

$$G_g(\text{cascaded}) \doteq Im_i\left[G_m(\text{cascaded}) \left(1 + i\sum_n \frac{G_{nqc}}{G_{nmc}}\right)\right]$$
$$\doteq G_m(\text{cascaded}) \sum_n \frac{G_{nqc}}{G_{nmc}} \quad (26)$$

The approximation of equation 26 is not a particularly useful form in general, although it may be used to advantage in special cases such as cascaded identical elements. It does serve to indicate that the relative amount of quadrature carrier tends to increase as the number of cascaded elements is increased.

In summary, it is concluded that for cascades of type-1 elements, arranged so that the output carrier phase of the last element is the proper value for maximum response of the type-3 element following (as is usually the case), the approximate over-all envelope transfer function is the product of the envelope transfer functions of the individual elements, each transfer function corresponding to the carrier component in phase with the element output for steady-state carrier input.

Envelope Transfer Functions of Representative Type-1 Elements

Capacitance-Resistance Coupling Circuit

The circuit of Fig. 2 is commonly used in coupling amplifier stages. The transfer function at actual frequency is

$$G(s) = \frac{Ts}{Ts+1}, \quad T = RC \quad (27)$$

The phase shift at carrier frequency is

$$\Phi_c = \tan^{-1}\frac{1}{\omega_c T} \quad (28)$$

The entire envelope transfer function is

$$\mathcal{G} = G(p+i\omega_c) = \frac{Tp(Tp+1)+\omega_c^2T^3+i\omega_c T}{(Tp+1)^2+\omega_c^2T^2} \quad (29)$$

Using equation 15, G_{mc} and G_{qc} are found to be

$$G_{mc} = (\cos\Phi_c)\frac{T^2p^2+Tp+\omega_c^2T^3+\omega_c T\tan\Phi_c}{T^2p^2+2Tp+1+\omega_c^2T^2}$$
$$= \frac{\omega_c T}{\sqrt{1+\omega_c^2T^2}}\,\frac{T^2p^2+Tp+1+\omega_c^2T^2}{T^2p^2+2Tp+1+\omega_c^2T^2} \quad (30A)$$

$$G_{qc} = (-\sin\Phi_c)\times \frac{T^2p^2+Tp+\omega_c^2T^2-\omega_c T\cot\Phi_c}{T^2p^2+2Tp+1+\omega_c^2T^2}$$
$$= \frac{-1}{\sqrt{1+\omega_c^2T^2}}\,\frac{Tp(Tp+1)}{T^2p^2+2Tp+1+\omega_c^2T^2} \quad (30B)$$

Letting

$$a = \omega_c T \quad (31)$$

G_{mc} and G_{qc} may be rewritten

$$G_{mc} = \frac{a}{\sqrt{1+a^2}}\,\frac{\dfrac{a^2}{1+a^2}\left(\dfrac{p}{\omega_c}\right)^2+\dfrac{a}{1+a^2}\left(\dfrac{p}{\omega_c}\right)+1}{\dfrac{a^2}{1+a^2}\left(\dfrac{p}{\omega_c}\right)^2+\dfrac{2a}{1+a^2}\left(\dfrac{p}{\omega_c}\right)+1} \quad (32A)$$

$$G_{qc} = \frac{-a}{(1+a^2)^{\frac{3}{2}}}\,\frac{\dfrac{p}{\omega_c}\left(a\dfrac{p}{\omega_c}+1\right)}{\dfrac{a^2}{1+a^2}\left(\dfrac{p}{\omega_c}\right)^2+\dfrac{2a}{1+a^2}\left(\dfrac{p}{\omega_c}\right)+1} \quad (32B)$$

Inspection of the quadratic polynomials in p which appear in equations 32(A) and 32(B) reveals that, regardless of the value of a, these quadratics have break frequencies $\geq\omega_c$ and damping factors ≤ 1. Thus if $\omega_m/\omega_c \leq 0.2$, the quadratics hardly affect the values of $G_{mc}(i\omega_m)$ and $|G_{qc}(i\omega_m)|$. Thus

$$\left|\frac{G_{qc}}{G_{mc}}\right| \doteq \frac{\dfrac{\omega_m}{\omega_c}\sqrt{1+a^2\dfrac{\omega_m^2}{\omega_c^2}}}{1+a^2}, \quad \frac{\omega_m}{\omega_c} \leq 0.15 \quad (33)$$

It is readily seen that $|G_{qc}/G_{mc}|$ decreases steadily as a increases from zero; setting $a = 0$, there results the upper bound

$$\left|\frac{G_{qc}}{G_{mc}}\right| \leq \frac{\omega_m}{\omega_c}, \quad \frac{\omega_m}{\omega_c} \leq 0.15 \quad (34)$$

which has been mentioned. Of course, practical coupling networks use large values of a which will greatly reduce

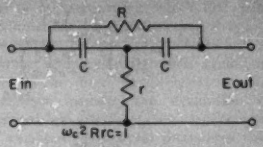


Fig. 5. Bridged-T equalizing network

$|G_{qc}/G_{mc}|$. If, for example, $a = 5$, and $\omega_m/\omega_c = 0.15$, then

$$\left|\frac{G_{qc}}{G_{mc}}\right| = \frac{0.15\sqrt{1.565}}{26} = \frac{0.15}{20.8} \quad (35)$$

As for G_{mc}, in the frequency band of interest the amplitude is essentially flat and the phase lag varies nearly linearly with frequency

$$|G_{mc}| \doteq a(1+a^2)^{-\frac{1}{2}}, \quad \frac{\omega_m}{\omega_c} \leq 0.15 \quad (36A)$$

$$/G_{mc} \doteq -a(1+a^2)^{-1}\frac{\omega_m}{\omega_c}, \quad \frac{\omega_m}{\omega_c} \leq 0.15 \quad (36B)$$

The condition of maximum phase lag occurs for $a = 1$ ($\Phi_c = 45$ degrees). For $\omega_m/\omega_c = 0.2$ and $a = 1$, the phase lag is about 1/10 radian or 6 degrees. Thus the presence of several of these coupling networks may contribute materially in deteriorating the phase margin of a system which has been designed without considering these effects.

Resistance-Capacitance Filter or Phase Shift Circuit

When the same procedure is used as before, the circuit of Fig. 3 gives the following equation

$$G(s) = \frac{1}{1+TS}, \quad T = RC \quad (37)$$

The phase shift at carrier frequency is

$$\Phi_c = -\tan^{-1}\omega_c T \quad (38)$$

The entire envelope transfer function is

$$\mathcal{G} = G(p+i\omega_c) = \frac{1+Tp-i\omega_c T}{(1+Tp)^2+\omega_c^2T^2} \quad (39)$$

Using equations 15 and 38 yields the following:

$$G_{mc} = \frac{1}{\sqrt{1+\omega_c^2T^2}}\,\frac{Tp+1+\omega_c^2T^2}{T^2p^2+2Tp+1+\omega_c^2T^2} \quad (40A)$$

$$G_{qc} = \frac{\omega_c T}{\sqrt{1+\omega_c^2T^2}}\,\frac{Tp}{T^2p^2+2Tp+1+\omega_c^2T^2} \quad (40B)$$

With $a = \omega_c t$ as before, these may be rewritten

$$G_{mc}=(1+a^2)^{-\frac{1}{2}}\frac{\frac{a}{1+a^2}\left(\frac{p}{\omega_c}\right)+1}{\frac{a^2}{1+a^2}\left(\frac{p}{\omega_c}\right)^2+\frac{2a}{1+a^2}\left(\frac{p}{\omega_c}\right)+1} \tag{41A}$$

$$G_{qc}=a^2(1+a^2)^{-\frac{1}{2}}\frac{\frac{p}{\omega_c}}{\frac{a^2}{1+a^2}\left(\frac{p}{\omega_c}\right)^2+\frac{2a}{1+a^2}\left(\frac{p}{\omega_c}\right)+1} \tag{41B}$$

Once again, the corner frequencies of the linear and the quadratic factor equal or exceed ω_c for any value of a, so that

$$\left|\frac{G_{qc}(i\omega_m)}{G_{mc}(i\omega_m)}\right|\doteq\frac{a^2}{1+a^2}\frac{\omega_m}{\omega_c},\ \frac{\omega_m}{\omega_c}\leq 0.15 \tag{42}$$

As for the previous circuit, the upper bound is ω_m/ω_c, in this case for $a\to\infty$. If the network is used for phase-shifting purposes with $a\geq 1$, the upper bound may be closely approached.

In this case also, G_{mc} is essentially flat in amplitude and linear in phase, with

$$|G_{mc}|\doteq(1+a^2)^{-\frac{1}{2}},\ \frac{\omega_m}{\omega_c}\leq 0.15 \tag{43A}$$

$$\underline{/G_{mc}}\doteq -a(1+a^2)^{-\frac{1}{2}}\frac{\omega_m}{\omega_c},\ \frac{\omega_m}{\omega_c}\leq<0.15 \tag{43B}$$

Again, the phase lag is greatest for $a=1$. Since it is quite likely that the value of a will be in the vicinity of unity for a phase shift circuit, the remarks made previously concerning the loss of phase margin apply here a fortiori.

Bridge Phase Shift Circuit

The bridge phase shift circuit shown in Fig. 4 is useful in varying carrier phase over a range of 180 degrees. Proceeding as before

$$G(s)=\frac{Ts-1}{Ts+1},\ \Phi_c=-\tan^{-1}\frac{2\omega_cT}{T^2\omega_c^2-1} \tag{44}$$

$$\mathcal{G}=G(p+i\omega_c)=\frac{T^2p^2-1+\omega_c^2T^2+2i\omega_cT}{(Tp+1)^2+\omega_c^2T^2} \tag{45}$$

so that

$$G_{mc}=\frac{T^2\omega_c^2-1}{T^2\omega_c^2+1}\frac{T^2p^2+\left(\frac{T^2\omega_c^2+1}{T^2\omega_c^2-1}\right)^2}{T^2p^2+2Tp+1+\omega_c^2T^2} \tag{46A}$$

$$G_{qc}=-\frac{2T\omega_c}{T^2\omega_c^2+1}\frac{T^2p^2}{T^2p^2+2Tp+1+\omega_c^2T^2} \tag{46B}$$

Rewriting with $a=\omega_cT$ yields

$$G_{mc}=\frac{\frac{a^2(a^2-1)}{(a^2+1)^2}\left(\frac{p}{\omega_c}\right)^2+1}{\frac{a^2}{1+a^2}\left(\frac{p}{\omega_c}\right)^2+\frac{2a}{1+a^2}\left(\frac{p}{\omega_c}\right)+1} \tag{47A}$$

$$G_{qc}=-\frac{2a^2}{(1+a^2)^2}\frac{\left(\frac{p}{\omega_c}\right)^2}{\frac{a^2}{1+a^2}\left(\frac{p}{\omega_c}\right)^2+\frac{2a}{1+a^2}\left(\frac{p}{\omega_c}\right)+1} \tag{47B}$$

Since the corner frequency of the factors involved equals or exceeds ω_c for any value of a

$$\left|\frac{G_{qc}(i\omega_m)}{G_{mc}(i\omega_m)}\right|\doteq\frac{2a^2}{(1+a^2)^2}\left(\frac{\omega_m}{\omega_c}\right)^2,\ \frac{\omega_m}{\omega_c}\leq 0.15 \tag{48}$$

Equation 48 has an upper bound, for $a=1.15$ approximately, of $0.565\ (\omega_m/\omega_c)^2$. This is considerably less than ω_m/ω_c, in the range of interest.

Again, G_{mc} is essentially flat in amplitude and linear in phase in the frequency range of interest, with

$$|G_{mc}|\doteq 1 \tag{49A}$$

$$\underline{/G_{mc}}\doteq -\frac{2a}{1+a^2}\frac{\omega_m}{\omega_c},\ \frac{\omega_m}{\omega_c}\leq 0.15 \tag{49B}$$

Here the phase lag is double that of either of the simple R/C (resistance-capacitance) configurations previously investigated. The value of $a=1$ again causes the greatest phase lag. For $\omega_m/\omega_c=0.15$ and $a=1$, the lag is about 0.15 radian or 8.5 degrees.

Bridged-T Equalizing Network

The networks previously investigated by the envelope transfer function technique have had only a second-order effect on the modulation envelope. This is not the case with the bridged-T circuit of Fig. 5, which is useful as an "a-c lead" network. It is of particular interest to compare the transfer function, obtained by using the narrow-bandwidth approximation, with the results obtained by the methods of this paper.

The transfer function $G(s)$ is readily found to be

$$G(s)=\frac{Rrc^2s^2+2rcs+1}{Rrc^2s^2+(2r+R)cs+1} \tag{50}$$

Letting

$$K^2=\frac{R}{r} \tag{51}$$

and substituting the resonance condition $\omega_c^2C^2rR=1$, $G(s)$ may be rewritten as

$$G(s)=\frac{2}{2+K^2}\frac{1+\frac{(s-j\omega_c)(s+j\omega_c)}{2s\omega_c}K}{1+\frac{(s-j\omega_c)(s+j\omega_c)}{2s\omega_c}\frac{2K}{2+K^2}} \tag{52}$$

If the signal bandwidth is assumed to be small compared with the carrier frequency, the "narrow-bandwidth" approximation is

$$\frac{s+j\omega_c}{2s}\doteq 1 \tag{53}$$

When equation 53 is used and $s-j\omega_c$ identified as p, equation 52 becomes

$$G_1(p)\doteq\frac{2}{2+K^2}\frac{1+\left(\frac{p}{\omega_c}\right)K}{1+\left(\frac{p}{\omega_c}\right)\frac{2K}{2+K^2}} \tag{54}$$

where $G_1(p)$ is the approximate signal transfer function given by the narrow-bandwidth approximation.

The envelope transfer function technique is next employed. The entire envelope transfer function is

$$\mathcal{G}=G(p+i\omega_c)=\frac{Rrc^2(p+i\omega_c)^2+2rc(p+i\omega_c)+1}{Rrc^2(p+i\omega_c)^2+(2r+R)c(p+i\omega_c)+1} \tag{55}$$

This may be simplified to

$$\mathcal{G}=\frac{\left(\frac{p}{\omega_c}\right)^2+\frac{2p}{\omega_cK}+2i\left(\frac{p}{\omega_c}+\frac{1}{K}\right)}{\left(\frac{p}{\omega_c}\right)^2+\frac{p}{\omega_cK}(2+K^2)+2i\left(\frac{p}{\omega_c}+\frac{1}{2K}(2+K^2)\right)} \tag{56}$$

Since $G(i\omega_c)=r/r+R/2=2/2+K^2$, then $\Phi_c=0$, and G_{mc} is given by the real part of $\mathcal{G}$ directly.

After simplification, G_{mc} may be written as

$$G_{mc}=\frac{\left(\frac{p}{\omega_c}\right)^4+\left(\frac{p}{\omega_c}\right)^2\left(\frac{4+K^2}{K}\right)+2\left(\frac{p}{\omega_c}\right)^2\left(\frac{2+3K^2}{K^2}\right)+2\left(\frac{p}{\omega_c}\right)\left(\frac{4+K^2}{K}\right)+\frac{2}{K^2}(2+K^2)}{\left(\frac{p}{\omega_c}\right)^4+2\left(\frac{p}{\omega_c}\right)^2\left(\frac{2+K^2}{K}\right)+\left(\frac{p}{\omega_c}\right)^2\left(\frac{K^4+8K^2+4}{K^2}\right)+4\left(\frac{p}{\omega_c}\right)\left(\frac{2+K^2}{K}\right)+\left(\frac{2+K^2}{K}\right)^2} \tag{57}$$

For particular numerical values of K, the biquadratic polynomials of equation 57 may be factored and the result compared with the narrow-bandwidth approximation of equation 54. It is more useful, however, to note that, if appreciable phase lead is desired in the region of the signal spectrum, K must be considerably larger than unity, as may be seen from the numerator of equation 54 (of course, this means the gain $2/2+K^2$ will be small). For large K, the biquadratic polynomials of equation 57 may be approximately factored into

$$G_{mc} \doteq \frac{\left(\frac{p}{\omega_c}+\frac{1}{K}\right)\left(\frac{p}{\omega_c}+K\right)\left(\left(\frac{p}{\omega_c}\right)^2+\frac{3}{K}\left(\frac{p}{\omega_c}\right)+2\right)}{\left(\frac{p}{\omega_c}+K\right)^2\left(\left(\frac{p}{\omega_c}\right)^2+\frac{2}{K}\left(\frac{p}{\omega_c}\right)+1\right)}$$

$$\doteq \frac{2}{K^2}\,\frac{\left(1+K\left(\frac{p}{\omega_c}\right)\right)\left(1+\frac{1.5}{K}\left(\frac{p}{\omega_c}\right)+\frac{1}{2}\left(\frac{p}{\omega_c}\right)^2\right)}{\left(1+\frac{1}{K}\left(\frac{p}{\omega_c}\right)\right)\left(1+\frac{2}{K}\left(\frac{p}{\omega_c}\right)+\left(\frac{p}{\omega_c}\right)^2\right)} \qquad K>>1 \quad (58)$$

It is clear that both equations 54 and 58 agree in the essential linear numerator factor which produces the lead. As far as the residual dynamic effect is concerned, the remaining factors do not appreciably affect the amplitude (since all corner frequencies exceed ω_c for $K>>1$) and the residual phase lags are

$$\underline{/G_1(i\omega_m)} \text{ residual} \doteq -\frac{2}{K}\frac{\omega_m}{\omega_c}, \quad K>>1, \quad \frac{\omega_m}{\omega_c} \leq 0.15 \quad (59A)$$

$$\underline{/G_{mc}(i\omega_m)} \text{ residual} \doteq -\frac{1.5}{K}\frac{\omega_m}{\omega_c}, \quad K>>1, \quad \frac{\omega_m}{\omega_c} \leq 0.15 \quad (59B)$$

which are essentially in agreement.

Conclusion

The envelope transfer function technique has been shown to permit convenient and exact analysis of linear a-c servo systems. Furthermore, through the use of approximations which are quite accurate, the transfer function of type-1 elements in cascade may be found as the product of individual transfer functions, just as in direct-signal systems. The technique permits the evaluation of residual lags previously omitted because of cumbersome analytic procedures, and it is expected that it will prove valuable in connection with the synthesis of compensating networks for a-c systems.

References

1. Stabilization of Carrier Frequency Servomechanisms. A. Sobczyk. *Journal*, Franklin Institute, Philadelphia, Pa., July, Aug., Sept., 1948.

2. Effects of Carrier Shifts on Derivative Networks for A-C Servomechanisms, George M. Ahura. *AIEE Transactions*, vol. 70, pt. I, 1951, pp. 612–18.

3. Transient Analysis of A-C Servomechanisms, S. S. L. Chang. *Ibid.*, vol. 74, pt. II, Mar. 1955, pp. 30–37.

Analogue Computer Methods Applied to Steady-State A-C System Problems

H. A. KAHLE
ASSOCIATE MEMBER AIEE

H. M. McCONNELL
MEMBER AIEE

THE GROWING complexity of aircraft a-c power systems has required more and more detailed analysis of the system performance under abnormal conditions, such as asymmetrical loading or asymmetrical faults. These calculations are used to predict the proper setting of the different protection functions and to assure proper co-ordination of primary and secondary protection in the distribution network. Another factor is concerned with the reduction of positive sequence voltage is applied to polyphase a-c motors. These motors are being applied in aircraft in increasing numbers.

Problems of this sort are not new to the industrial power system engineer. They become relatively more severe in the aircraft a-c system, in that the voltage regulation and system protection functions are interconnected and interdependent. For example, the voltage regulators in present use derive their primary power from the terminals of the a-c system. To design regulators for the desired performance under abnormal conditions it is necessary to know the behavior of the system voltages.

The voltage of an aircraft a-c system is regulated according to the output of a 3-phase full-wave bridge rectifier as the sensing element. Recently, regulators have been employed which select the highest of the three line-to-neutral voltages and regulate that quantity. In either case, the line-to-neutral voltages of the a-c system under unbalanced conditions must be calculated, both in magnitude and in phase, in order to determine the regulated level. In the case of 3-phase full-wave sensing such information allows the computation of overvoltage during steady faults, and in the second case, the calculation is necessary for to determining which phase is regulated.

Digital methods of calculation have been used for investigations of this type.[1,2] The digital techniques are well suited to this work. This paper presents an alternate method of calculation for steady-state a-c system performance which makes use of the electronic differential analyzer.

Paper 56-766, recommended by the AIEE Air Transportation Committee and Computing Devices Committee of the Communication and Electronics Division, and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Summer and Pacific General Meeting, San Francisco, Calif., June 25–29, 1956. Manuscript submitted January 30, 1956; made available for printing May 22, 1956.

H. A. Kahle and H. M. McConnell are with Jack & Heintz, Inc., Cleveland, Ohio.

Analogue Computation Methods

The theory of symmetrical components is used to develop an equivalent circuit for the unbalanced condition under investigation. Such circuits are in common use in industrial a-c system work and are thoroughly presented in many textbooks and handbooks.[3,4] The objective of the calculation is to determine the line-to-neutral voltages, in magnitude and phase, as a function of fault current, with parameters, for example, as external balanced load or transmission line impedances.

The equivalent circuit of line-to-ground resistive system fault is shown in Fig. 1. By combining the parallel impedances of the machine, transmission line, and load into a new equivalent impedance, the circuit diagram can be reduced to that shown in Fig. 2. This type of circuit is normally described by a set of complex algebraic equations. In this form, however, the problem cannot be adapted for solution on analogue computers. The complex algebra has proved very helpful in electrical engineering problems. It describes electrical quantities as phasors in the complex plane with fixed co-ordinates. The symbolic arithmetic operation treats the phasors as being stationary in the complex plane. If it is remembered that these phasors actually are rotating, and the electrical quantities at any time are obtained by the projection of the phasors on the imaginary axis, the set of equations may be converted into a form in which time is an independent variable. In this form, the problem will be suitable for solution on an analogue computer.

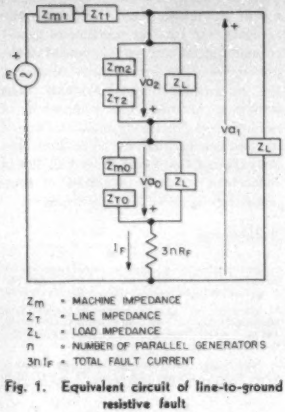


Fig. 1. Equivalent circuit of line-to-ground resistive fault

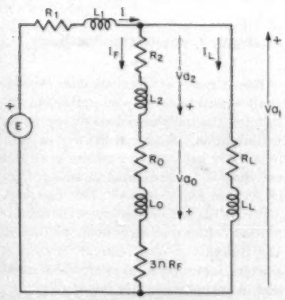


Fig. 2. Reduced equivalent circuit of line-to-ground resistive fault

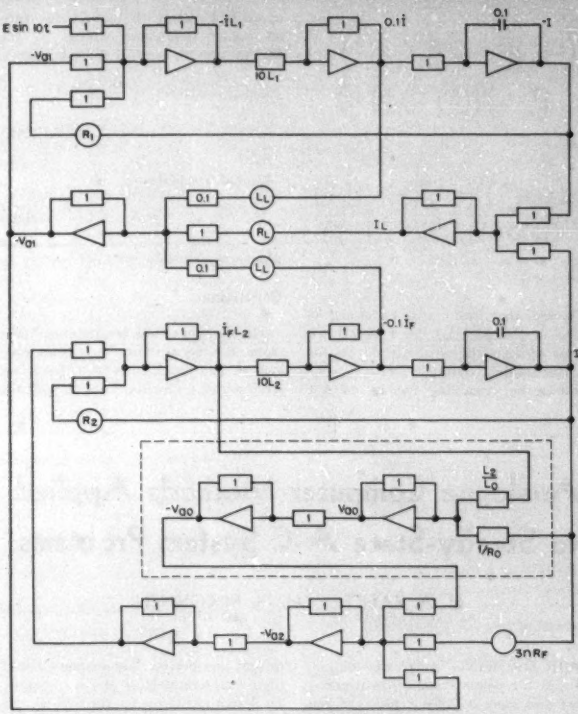


Fig. 3. Computer representation for line-to-ground resistive fault

The following equations, all variables being a function of time, describe the system as shown in Fig. 2.

$$E - V_{a1} = IR_1 + \frac{d}{dt}(IL_1)$$

$$-V_{a2} = I_F R_2 + \frac{d}{dt}(I_F L_2)$$

$$V_{a1} = I_L R_L + \frac{d}{dt}(I_L L_L)$$

$$I_F = I - I_L - V_{a2} = V_{a1} + V_{a0} - I_F(3nR_F)$$

$$-V_{a0} = I_F R_0 + \frac{d}{dt}(I_F L_0)$$

The system under study was a power system at 400 cycles per second. The problem therefore could not be solved on the computer in real time, because of the restricted frequency range of the recording equipment. For computer operation the time has been changed to $\omega = 10$. The Laplace transform time scale theorem may be used to transform the real time equations into computer time equations. However, the use of 10 radians per second allows the conversion of reactance values into inductance values by merely shifting the decimal point by one place to the left. From these equations, a computer block diagram can readily be drawn; see Fig. 3. It gives the fault current, positive, and zero-sequence voltages as functions of exciting electromotive force with the fault resistance as parameter.

The phase voltages present in the system are composed of the voltages available in the computer block diagram, rotated by 120 or 240 degrees. A unity phasor a has been adopted for this phase rotation by means of the theory of symmetrical components. Shifting phasors by 120 degrees on a computer normally requires a servo resolver. Since all variables are given here as a function of time, the multiplication by the unity phasor a can be written as follows.

$$aV = Ve^{j2\pi/3} = -0.5V + j0.866V$$

$$aV = -0.5V - 0.866\omega \int V dt$$

This equation indicates that the phase shift of 120 degrees can be performed by summation and integration, two operations suited to the analogue computer. Fig. 4 shows how the positive- and negative-sequence voltages are rotated by 120 and 240 degrees respectively. The synthesis of the resulting voltages to the actual phase voltages in the system is described by the following three equations:

$$V_a = V_{a1} + V_{a2} + V_{a0}$$

$$V_b = a^2 V_{a1} + a V_{a2} + V_{a0}$$

$$V_c = a V_{a1} + a^2 V_{a2} + V_{a0}$$

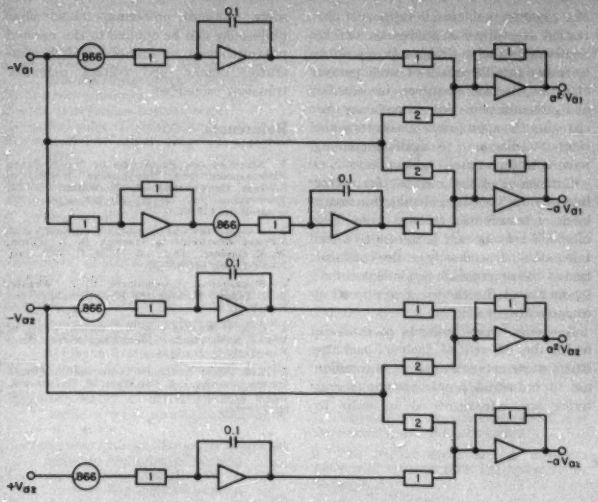


Fig. 4. Phase rotation by 120 and 240 degrees

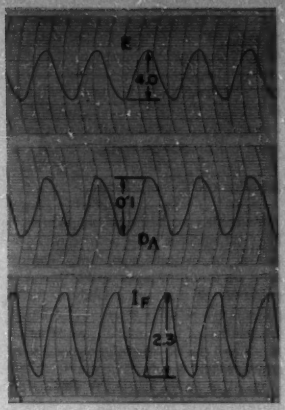


Fig. 6. E, V_a, and I_F at line-to-line solid short circuit

Fig. 5 shows the computer diagram.

As described, the voltage regulator connected to the generator tends to hold one of the phase voltages at rated value. Since all three voltages are available on the computer, the highest one can easily be selected, The computer input voltage E must be adjusted to keep the highest of the 3-phase voltages at one per unit for any value of fault resistance.

The computer diagram as shown in Fig. 5 for the solution of line-to-ground system faults requires 26 amplifiers. The system parameters can be adjusted and varied by changing the scale factor potentiometers. Plotting of voltages and fault current on a multichannel recorder permits direct reading of magnitude and phase shift.

Line-to-line resistive fault is another type of fault which often must be investigated in system performance. In the equivalent block diagram for this type of fault, the zero-sequence impedance has to be short-circuited in Figs. 1 and 2. The

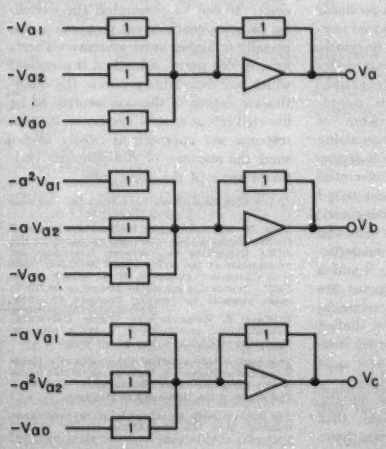


Fig. 5 (left). Synthesis of line-to-neutral voltages

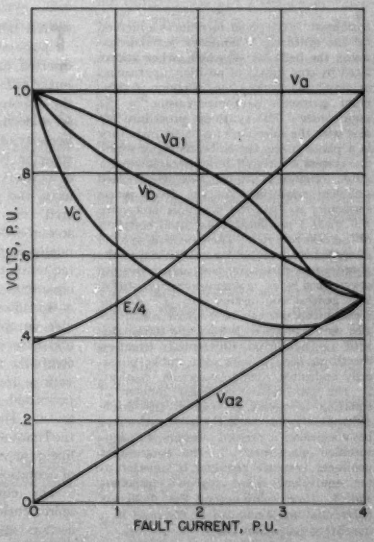


Fig. 7 (right). Line-to-line resistive fault

sign of the negative-sequence voltage reverses. Thus the block diagram can easily be changed to represent line-to-line resistive faults by taking the two amplifiers out which are shown in dashed lines in Fig. 3. The actual fault current becomes $\sqrt{3}\ I_F$ and the fault resistance is equal to R_F.

Results and Conclusions

A sample calculation for single line-to-line fault has been made both numerically and on the analogue computer. The results are found to agree within the precision of the differential analyzer. Representative recorder traces taken during the solution of this problem are shown in Fig. 6, and the curves giving the complete results in Fig. 7.

This use of the electronic differential analyzer allows the solution of a variety of a-c system problems, in support of protection co-ordination studies as well as regulator and relay design. It is possible to investigate the effect of such parameters as machine sequence impedances and transmission line impedances by changing the appropriate constants in the computer diagram. A significant saving in computation time is achieved.

Extensions of the method, for taking into account the effect of induction motor loads of transformer-rectifier units, both of which act as phase balancers to a certain extent, depend only on the construction of the appropriate equivalent circuit by means of the theory of symmetrical components.

A recent paper[5] presents methods in which the differential analyzer and the steady-state network analyzer in common use in industrial power system laboratories are interconnected in order to solve transient problems. The same philosophy can be applied to the method presented here, by simulation of the generator, exciter, and voltage regulator transient properties.

References

1. Methods for Prediction of Steady-State Performance for Unbalanced Regulated 3-Phase Generators, B. J. Wilson. *AIEE Transactions*, vol. 72, pt. II, 1953 (Jan. 1954 section), pp. 413–22.

2. Calculations on Voltage Unbalance for 3-Phase Synchronous Systems, B. J. Wilson, W. K. Gardner. *Ibid.*, vol. 73, pt. II, 1954 (Jan. 1955 section), pp. 426–37.

3. Symmetrical Components, C. F. Wagner, R. D. Evans. McGraw-Hill Book Company, Inc., New York, N. Y., 1933.

4. Circuit Analysis of A-C Power Systems, Vol. I, Edith Clarke. John Wiley & Sons, Inc., New York, N. Y., 1934.

5. The Use of Analogous Computers in Power System Studies, J. E. Van Ness, W. C. Peterson. *AIEE Transactions*, vol. 75, pt. III, Apr. 1956, pp. 238–42.

A Phase-Space Method for the Synthesis of Nonlinear Servomechanisms

RICHARD E. KUBA
ASSOCIATE MEMBER AIEE

LOUIS F. KAZDA
MEMBER AIEE

Synopsis: This paper introduces a method for the synthesis of nonlinear servomechanisms the behavior of which, when stimulated by step inputs of position, approaches as closely as desired the behavior of optimum contactor servomechanisms of the same order. The synthesis procedure begins with the selection of a phase trajectory in a phase space, the co-ordinates of which are system error and derivatives of system error. Theorems are proposed which place sufficient conditions on the given phase trajectory so that the synthesis procedure will yield a nonlinear differential equation of the desired type. The synthesis process is carried out in detail for nonlinear systems of order two, three, and four, and the results are verified by an analogue-computer study. The general conclusions are: 1. A synthesis procedure can be evolved which begins with a phase trajectory and terminates with a nonlinear differential equation describing the behavior of a stable, physically realizable, nonlinear servomechanism; 2. the synthesized nonlinear differential equation is closely related to the linear differential equation of the same order and has coefficients expressed in terms of design-specified quantities; 3. the synthesized nonlinear system's response is superior to the equivalent linear system's response; and 4. time solutions of the nonlinear differential equations can be obtained in a number of important cases.

THE BEHAVIOR of a control system possessing nonlinearities, inherent or inserted, can often be described by a nonlinear differential equation. Frequently, the response characteristics of a nonlinear servomechanism having an inserted nonlinearity of the proper type will be greatly superior to the response characteristics of an equivalent linear servo. McDonald,[1] using the phase-plane approach, considered the response characteristics of second-order servo systems containing nonlinear damping and nonlinear restoring terms respectively in their differential equations. Lewis[2] designed and tested a second-order servo incorporating variable damping. In a paper by Caldwell and Rideout,[3] the results of an analogue-computer study are presented in which various forms of variable damping are permitted in a second-order servomechanism. Howe and Rauch[4] have studied the characteristics of a second-order, nonlinear servo incorporating a linear zone of operation for small system errors.

The conclusions arrived at in the aforementioned papers indicate that certain variably damped servos have significant advantages over the equivalent linear servo.

The ultimate objective of introducing nonlinearities into a servo system is to produce an optimum response characteristic exemplified by the equivalent contactor servomechanism of same order. Bogner and Kazda[5] have shown that an nth-order contactor servomechanism exhibited minimum response time when subjected to $(n-1)$ switching criteria. However, because of the complexity of the switching criteria of minimum response time contactor servos, the control circuitry needed to accomplish the switching program reaches vast proportions, especially in higher order systems. Therefore, in this paper, a method is proposed which will determine precisely the quantitative nature of the nonlinearity to be inserted into a control system so that its response will approach as closely as desired the response of the optimum contactor servo of the same order.

Paper 56-455, recommended by the AIEE Feedback Control Systems Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE-ASME Nonlinear Controls Conference, Princeton, N. J., March 26–28, 1956. Manuscript submitted December 28, 1955; made available for printing February 14, 1956.

Richard E. Kuba is with Wayne University, Detroit, Mich., and Louis F. Kazda is with the University of Michigan, Ann Arbor, Mich.

This work contains part of the results of a thesis submitted by Richard E. Kuba in partial fulfillment of the requirements for the degree of Doctor of Philosophy at the University of Michigan.

The authors wish to acknowledge the assistance received from the engineering faculty at the University of Michigan.

Nomenclature

A_p = real constants
b_p = real coefficients in the differential equation describing system dynamics
B, C_p = real constants
E = servomechanism position error
E_1 = initial value of position error
E_{1M} = maximum initial value of position error
F = coefficient of viscous friction
$F_1(\ )$ = functional notation
$F_2(\ )$ = functional notation
$g(\ )$ = function of
G_h = amplifier gain coefficient
$h(\ )$ = function of
J = moment of inertia of motor and load
K_A = real constant
S = contactor servo switching constant
t = real time
T_A = time delay of energy storage device
T_{c2} = response time of second-order contactor servo to step input of position
T_{2n} = response time of the equivalent nth class of nonlinear second-order servo
T_{c2M} = response time of second-order contactor servo to maximum step input of position
T_{c3} = response time of third-order contactor servo to step input of position
T_{c4} = response time of fourth-order contactor servo to step input of position
$T_{1,2}$ = time between points 1 and 2 in the phase space
T_{n2} = response time of second-order class-2 nonlinear servo to step input of position
T_{n3} = response time of third-order class-2, nonlinear servo to step input of position
T_{n3M} = response time of third-order class-2, nonlinear servo to maximum step input of position
T_{n4} = response time of fourth-order class-2, nonlinear servo to step input of position
T_{n4M} = response time of fourth-order class-2, nonlinear servo to maximum step input of position
$u = E/E_1$ = analogue-computer variable
u_p = general phase-space co-ordinates corresponding to d^pE/dt^p
u_{s0} = co-ordinates of point lying on the u_s axis
$x = E$
$x_1 = E_1$
$x_2 = E_2$
$y = E'$
$z = E''$
$w = E'''$
$\zeta(\ )$ = Zeta function of
α = real positive constant
$\Gamma(\ )$ = Gamma function of
$\delta(E/E_1) = +1$ for $(^1/_4 < E/E_1 \leq 1)$, $\delta(E/E_1) = -1$ for $(0 \leq E/E_1 < ^1/_4)$
θ_i = servomechanism input variable
θ_o = servomechanism output variable
$\varphi(x)$ = function of x

Synthesis Approach

It is well known that analytical solutions to nonlinear differential equations can rarely be obtained. Thus, the analyst is unable to predict, analytically, how changes in his system parameters will affect the system response.

On the other hand, the synthesis approach to the problem begins with the response characteristics which are desired and terminates with a nonlinear ordinary differential equation exhibiting the given response characteristics as its solution. Moreover, the synthesized, nonlinear differential equation can be expressed so that its constants are given in terms of the design parameters specifying the maximum system initial error and the response time corresponding to the maximum initial error. Once the differential equation has been synthesized, the designer produces building blocks which, when properly connected together, will yield a control system the behavior of which is expressed by the synthesized differential equation.

The synthesis process must contain three fundamental properties:

1. The desired response characteristics of the control system must be known.
2. The desired form of the nonlinear differential equation to be synthesized must be specified.
3. Sufficient conditions must be known so that it can be determined whether or not the specified differential equation can be made to exhibit the desired response characteristics.

In this paper, the desired system response characteristics are obtained by approximating the phase trajectory of the optimum contactor servomechanism (when subjected to a step input of position) by a simple continuous curve. The desired form of the nonlinear differential equation will be

$$\frac{d^pE}{dt^p} + h(E)\frac{d^{p-1}E}{dt^{p-1}} + b_{p-2}\frac{d^{p-2}E}{dt^{p-2}} + \ldots + b_1E' + b_0E = 0, \ (p \geq 2) \quad (1)$$

Theorems are presented which establish sufficient conditions on a phase trajectory to render it suitable for the synthesis of equation 1.

Synthesis of Second-Order Systems

Theorem I and corollary specify sufficient criteria for the synthesis of a second-order nonlinear servomechanism.

Theorem I: If $\alpha y = \varphi(x)$, where $y = x'$, is continuous in the region $(x_1 \geq x \geq x_2)$ with $\varphi(x_1) = 0$ and has a continuous first derivative (dy/dx) in the region

$(x_1 > x > x_2)$

then the real constant a_1 and the function $f(x)$ can be uniquely determined such that the differential equation

$x'' + f(x)x' + a_1x = 0$ has $ay = \varphi(x)$

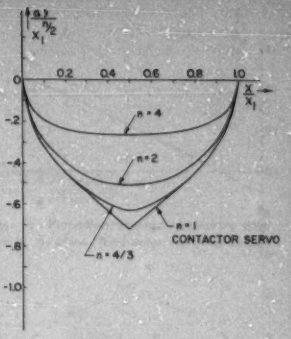


Fig. 1. Phase trajectories for various values of n

as its phase trajectory in the interval $(x_1 \geq x \geq x_2)$, with $f(x)$ a continuous function, except perhaps at points where $y = 0$, provided

$$\lim_{\substack{x \to x_1 \\ y \to 0}} [yf(x)] = 0$$

and

$$\lim_{\substack{x \to x_1 \\ y \to 0}} [ydy/dx]$$

exists and is nonzero.

Corollary: The real constant a_1 and the function $f(x)$ are given by $a_1 = g(x_1, x_2, T_{1,2})$ and $f(x) = h(x, x_1, x_2, T_{1,2})$, where $T_{1,2}$ is the absolute value of time elapsed from $x = x_1$ to $x = x_2$ provided

$$\left| \int_{x_1}^{x_2} \frac{dx}{y} \right|$$

exists and is nonzero.

Consider now the phase trajectories of the optimum second-order contactor servo for a step input of position, as illustrated in Fig. 1. The phase-plane co-ordinates are taken as system error and error rate multiplied by the constant α which gives both co-ordinates the same physical units. The problem of the synthesizer is to select a curve which will closely approximate the optimum phase trajectory in the region of interest. When this is done he is then able to use the criteria of theorem I and its corollary to test his selection.

A choice which will yield useful results is the family of trajectories given by the equations

$$y = -\frac{1}{\alpha}\sqrt{\left(\frac{x_1}{2}\right)^n - \left(x - \frac{x_1}{2}\right)^n} \quad (1 \leq n < \infty) \quad (2)$$
$$(x_1 \geq x \geq x_1/2)$$

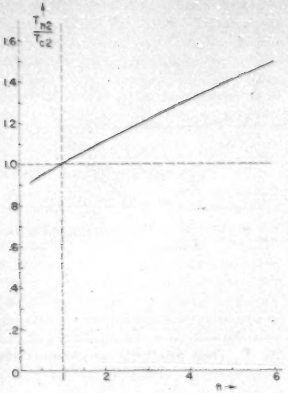


Fig. 2. Plot of T_{2n}/T_{c2} versus n

$$y=-\frac{1}{\alpha}\sqrt{\left(\frac{x_1}{2}\right)^n-\left(\frac{x_1}{2}-x\right)^n} \qquad \begin{matrix}(1\leq n<\infty)\\(x_1/2\geq x\geq 0)\end{matrix} \quad (3)$$

and is shown in Fig. 1 for various values of the parameter n. Theorem I is now applied to determine a_1 and $f(x)$ in the nonlinear differential equation

$$x''+f(x)x'+a_1x=0 \quad (4)$$

$$a_1=-\frac{1}{x_1}\lim_{\substack{x\to x_1\\ y\to 0}}\left(y\frac{dy}{dx}\right)=\frac{nx_1^{n-2}}{2^n\alpha^2} \quad (5)$$

Equating the acceleration of the nonlinear system at $t=0$ to the maximum acceleration of the optimum contactor system yields the equivalence relationship

$$a_1=\frac{4}{T_{c2}^2} \quad (6)$$

Therefore

$$\alpha=\sqrt{n}\,T_{c2}x_1^{n/2-1}/2^{n/2+1} \quad (7)$$

and

$$f(x)=-\frac{1}{y}\left(y\frac{dy}{dx}+a_1x\right)$$

$$f(x)=\frac{2\sqrt{n}}{T_{c2}}\left(\frac{\frac{x}{x_1}-\left|2\frac{x}{x_1}-1\right|^{n-1}\delta(x)}{\sqrt{1-\left|2\frac{x}{x_1}-1\right|^n}}\right) \qquad (0\leq x/x_1\leq 1) \quad (8)$$

Thus, the form of the general, synthesized second-order nonlinear differential equation is given in terms of system error and its derivatives by

$$E''+\frac{2\sqrt{n}}{T_{c2}}\left[\frac{\frac{E}{E_1}-\left|2\frac{E}{E_1}-1\right|^{n-1}\delta\left(\frac{E}{E_1}\right)}{\sqrt{1-\left|2\frac{E}{E_1}-1\right|^n}}\right]E'+\frac{4}{T_{c2}^2}E=0 \qquad \begin{matrix}(0\leq E\leq E_1)\\(1\leq n<\infty)\end{matrix} \quad (9)$$

The application of the corollary of theorem I will produce a relationship between T_{c2} and T_{2n}. Because of symmetry, the total response time for a step input of position will be just twice the time required to traverse the phase trajectory from $x=x_1$ to $x=x_1/2$

$$T_{2n}=2\int_{x_1}^{x_1/2}\frac{dx}{y}=2^{2/n-1}\sqrt{n}\times\frac{\left[\Gamma\left(\frac{n+1}{n}\right)\right]^2}{\Gamma\left(\frac{n+2}{n}\right)}T_{c2} \quad (10)$$

Equation 10 is plotted in Fig. 2, while the variable damping function is presented in Fig. 3 where $u=E/E_1$. Note particularly that the class-2 ($n=2$) nonlinear servo has a response time only 11 per cent greater than the response time of the optimum contactor system.

Time Solutions

Time solutions of equation 9 may be obtained for values of n equal to 4, 2, 4/3, and 1. Since

$$\int_0^t dt=\int_{x_1}^x\frac{dx}{y} \quad (11)$$

these solutions are

$$n=4,\ E(t)=\frac{E_1}{2}\left[1+cn\left(\frac{2\sqrt{2}t}{T_{c2}}\middle|\tfrac{1}{2}\right)\right] \qquad (0\leq t\leq 1.311T_{c2}) \quad (12)$$

Here, the elliptic cosine is expressed in the notation of Milne-Thompson[6]

$$n=2,\ E(t)=\frac{E_1}{2}\left(1+\cos\frac{2\sqrt{2}t}{T_{c2}}\right) \qquad (0\leq t\leq 1.111T_{c2}) \quad (13)$$

$$n=4/3,\ \frac{t}{T_{c2}}=0.27981cn^{-1}\times[(2E/E_1-1)^{1/3}|\tfrac{1}{2}]+1.22473\zeta\{cn^{-1}[(2E/E_1-1)^{1/3}|\tfrac{1}{2}]|\tfrac{1}{2}\}\ (0\leq t\leq 0.51875T_{c2}) \quad (14)$$

Here, the ζ function is expressed in the notation of Milne-Thompson[6]

$$n=1,\ E=E_1[1-2(t/T_{2c})^2] \qquad (0\leq t\leq T_{2c}/2) \quad (15)$$

$$E=2E_1(1-t/T_{c2})^2 \qquad (T_{c2}/2\leq t\leq T_{c2}) \quad (16)$$

Fig. 4 shows these time solutions together with their first and second derivatives.

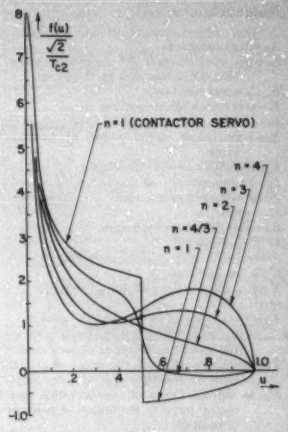


Fig. 3. Plot of the variable damping function for the synthesized second-order nonlinear servomechanism

Class-2 Second-Order Nonlinear Servo

Because of the simplicity of its time solutions and because its damping function always remains positive, the authors believe the $n=2$ nonlinear servo to be of special interest. The nonlinear differential equation of this system is given by

$$E''+\frac{\pi}{2T_{2n}}\sqrt{\frac{E_1}{E}-1}\,E'+\frac{\pi^2}{2T_{2n}^2}E=0 \quad (17)$$

where

$$T_{2n}=1.111T_{c2} \quad (18)$$

Equation 17 is not stated in the most convenient form for use in the construction of the physical servo system because the time T_{2n} is not a true constant but depends on the magnitude of the initial error E_1. Moreover, to make the system response symmetrical for positive and negative errors, E is replaced by $|E|$ in the variable damping coefficient of equation 17. Therefore, the practical form of the system's differential equation will read

$$E''+\frac{\sqrt{2}}{T_{c1M}}\sqrt{\frac{E_{1M}}{E_1}\sqrt{\frac{E_1}{|E|}}-1}\,E'+\frac{4E_{1M}}{T_{c1M}^2E_1}E=0 \quad (19)$$

having the time solution

$$E(t)=\frac{E_1}{2}\left[1+\cos\left(\frac{2}{T_{c1M}}\sqrt{\frac{2E_{1M}}{E_1}}t\right)\right] \qquad \left(0\leq t\leq 1.111\sqrt{\frac{E_1}{E_{1M}}}T_{c1M}\right) \quad (20)$$

Fig. 5 presents the variable damping co-

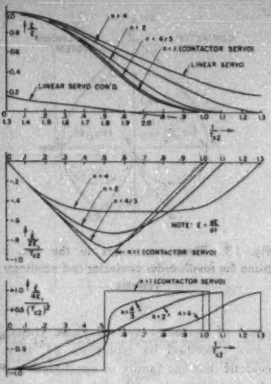


Fig. 4. Time solutions for various values of n for a second-order system

efficient of equation 19 plotted against E/E_{1M}.

It is evident that, since the nonlinear damping function of equation 19 and the coefficient $4E_{1M}/T_{c1M}{}^2E_1$ theoretically approach positive infinity for zero values of initial error, any practical control system the behavior of which is specified by equation 19 must terminate these quantities in positive finite values at zero initial error. Hence, the practical nonlinear servo will have an automatically built-in stable, linear mode of operation in the near vicinity of the origin of its phase space. Furthermore, the theorems of Liapounoff[7] and Bendixson[8] assure the stability of the singular point and the nonexistence of limit cycles. A block diagram of this

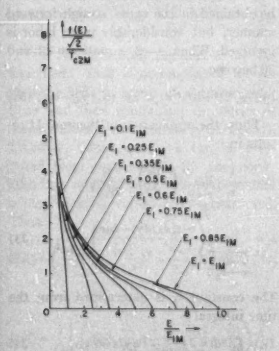


Fig. 5. Plot of the nonlinear damping function of equation 19

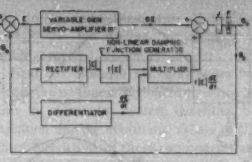


Fig. 6. Block diagram of the synthesized second-order nonlinear servomechanism

nonlinear servo is illustrated in Fig. 6. The equation describing the behavior of this servo is given by

$$E''+\frac{F+f(|E|)}{J}E'+\frac{G}{J}E=\theta_i''+\frac{F}{J}\theta_i' \qquad (21)$$

For inputs where, $d^q\theta_i/dt^q=0$, $(q\geq 1)$, $(t>0)$, and upon comparing equations 21 and 19, the variable-gain amplifier coefficient G and the nonlinear function generator's equation $f(|E|)$ are obtained as

$$G\cong\frac{4JE_{1M}}{T_{c1M}{}^2E_1} \qquad (22)$$

$$f(|E|)\cong\frac{\sqrt{2}J}{T_{2M}}\sqrt{\frac{E_{1M}}{E_1}}\sqrt{\frac{E_1}{|E|}}-1-F \qquad (23)$$

Equations 22 and 23 are written with approximately equal signs because at zero initial error both of these quantities must approach finite constants in a physically realizable system.

The system designer is now in a position to use his ingenuity for producing physical building blocks meeting the requirements of equations 22 and 23. Interconnection of these building blocks will result in a system described by equation 19 and having a theoretical response time which exceeds the optimum contactor servo response time by only 11 per cent.

Higher Order Systems

The starting point for the synthesis of high-order nonlinear servomechanisms will be the projection of the phase trajectory of the corresponding contactor control system described by the equation

$$\frac{d^pE}{dt^p}=\pm S \qquad (p=2, 3, 4, \ldots) \quad (24)$$

onto the u_{p-1}, u_{p-2} plane. These projected curves consist of a series of parabolas joined together, as illustrated in Figs. 7 and 8 for $p=3$ and $p=4$ respectively. The response of this form of contactor system will be considered optimum in the subsequent discussion. The problem of the synthesizer is to approximate the trajectory of the contactor system

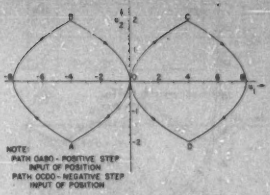


Fig. 7. Projection of phase trajectory onto the u_1, u_2 plane, third-order contactor system

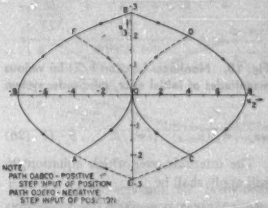


Fig. 8. Projection of phase trajectory onto the u_1, u_2 plane, fourth-order contactor system

given by equation 24 with a curve or set of curves such that the function $h(E)$ and the constants b_0, b_1, . . . , and b_{p-2} in the nonlinear differential equation

$$\frac{d^pE}{dt^p}+h(E)\frac{d^{p-1}E}{dt^{p-1}}+b_{p-2}\frac{d^{p-2}E}{dt^{p-2}}+\ldots+b_1E'+b_0E=0 \qquad (25)$$

can be uniquely determined when the system is subjected to a step input of position.

Generalization of Theorem I

Theorem II: Let the projections of the phase trajectory of the nonlinear system on the αu_{p-1}, u_{p-2} plane be composed of $(p-1)$ segments where the equation of the ith segment is given by

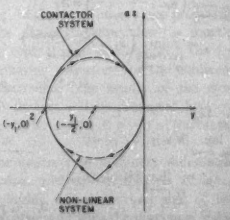


Fig. 9. Phase trajectories in the αx, y plane for third-order contactor and nonlinear systems

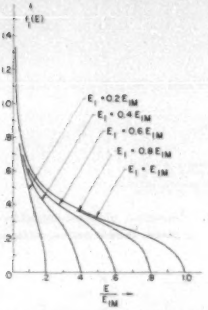


Fig. 10. Nonlinear function $h_i(E)$ for various magnitudes of initial error, third-order system

$$\alpha u_{p-1} = g_i(u_{p-2}) \quad (i=1, 2, \ldots, p-1) \quad (26)$$

The intervals over which equation 26 will apply shall be

$$C_{i-1} \leq u_{p-2} \leq C_i \quad (i=1, 2, \ldots, p-1) \quad (27)$$

Allow the following premises to be true:

1. Let the phase trajectory begin at a point on the u_0 axis and end at the origin of the phase space.

2. Let the projections of the phase trajectory onto the planes u_i, u_0, where $i=1, 2, \ldots, p-1$, describe curves such that along each of these curves (in their respective planes) the u_i are continuous functions of u_0 over the entire span of the trajectory including the end points.

3. Let the functions $g_i(u_{p-2})$, $(i=1, 2, \ldots, p-1)$ be continuous over their closed intervals of definition.

4. Let $[dg_i(u_{p-2})]/(du_{p-2})$, $(i=1, 2, \ldots, p-1)$ be continuous functions over the open intervals of definition of $g_i(u_{p-2})$.

5. Let $g_i(C_{i-1}) = 0 = g_i(C_i)$, $(i=1, 2, \ldots, p-1)$.

6. Let

$$\lim_{\substack{u_{p-2}\to C_0 \\ u_{p-1}\to 0}} \left(u_{p-1} \frac{dg_1(u_{p-2})}{du_{p-2}} \right) = A_0$$

$$\lim_{\substack{u_{p-2}\to C_i \\ u_{p-1}\to 0}} \left(u_{p-1} \frac{dg_i(u_{p-2})}{du_{p-2}} \right) = A_i = \lim_{\substack{u_{p-2}\to C_i \\ u_{p-1}\to 0}} \left(u_{p-1} \frac{dg_{i-1}(u_{p-2})}{du_{p-2}} \right) \quad (i=1, 2, \ldots, p-2)$$

7. Let

$$\lim_{u_{p-1}\to 0} [h(u_0) g_i(C_{i-1})] = 0 = \lim_{u_{p-1}\to 0} [h(u_0) g_i(C_i)] \quad (i=1, 2, \ldots, p-1)$$

If these premises are true, then the constants $b_0, b_1, \ldots, b_{p-2}$ and the function $h(E)$ of equation 25 can be uniquely determined such that equation 25 exhibits a phase trajectory the projection of which onto the u_{p-1}, u_{p-2} plane is given by the $(p-1)$ segments defined in the hypothesis, and which has a function $h(u_0) = h(E)$ which is continuous over the entire span of the trajectory except, perhaps, at points where u_{p-1} equals zero.

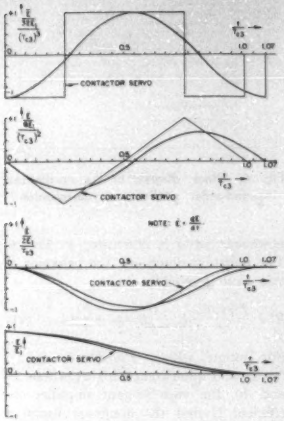


Fig. 11. Time solutions for third-order contactor and nonlinear servomechanisms

Synthesis of a Third-Order Nonlinear Servomechanism

The nonlinear differential equation to be synthesized will have the form

$$E''' + h(E)E'' + b_1E' + b_0E = 0 \quad (28)$$

The problem is to determine b_0, b_1, and $h(E)$ so that the time response of the nonlinear system approaches the time response of the optimum third-order contactor servo.

Because of the parabolic nature exhibited by projections on the u_{p-1}, u_{p-2}

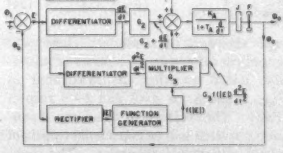


Fig. 12. Block diagram of third-order nonlinear servo

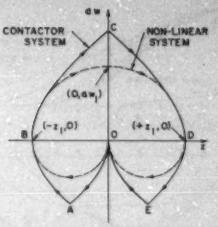


Fig. 13. Phase trajectories in the αw, z plane for fourth-order contactor and nonlinear systems

plane of contactor system phase trajectories described by equations 24,[1] it is evident that the family of approximating phase trajectories selected for second-order systems can be easily adapted to approximate these parabolic curves. Fig. 9 presents the parabolic phase trajectory projections of the third-order contactor servo together with a typical trajectory of the nonlinear system defined by

$$(y+y_1/2)^n + \alpha^2 z^2 = (y_1/2)^n \quad (0 \geq y \geq -y_1/2) \quad (29)$$

$$(-y-y_1/2)^n + \alpha^2 z^2 = (y_1/2)^n \quad (-y_1/2 \geq y \geq -y_1) \quad (30)$$

where n is allowed to take on values in the range $(1 \leq n < \infty)$. Clearly, as n approaches unity, the nonlinear system phase trajectories move closer to those of the parabolic system. Because the nonlinear system of class-2 $(n=2)$ has a response time only a few per cent greater than the response time of the equivalent contactor system and, because of the simplicity of its time solutions, the derivation will proceed on the basis that $n=2$. The results for other values of n can be obtained in the same straightforward manner, but considerably more labor is involved. When $n=2$, equations 29 and 30 become

$$(y+y_1/2)^2 + \alpha^2 z^2 = (y_1/2)^2 \quad (0 \geq y \geq -y_1) \quad (31)$$

Thus, the application of theorem II results in

$$b_0 = -\frac{1}{x_1} \lim_{\substack{y\to 0 \\ z\to 0}} \left(z \frac{ds}{dy} \right) = \frac{y_1}{2x_1\alpha^2} \quad (32)$$

$$b_1 = \lim_{\substack{y\to -y_1 \\ z\to 0 \\ x\to x_1/2}} \left(\frac{-s(ds/dy) - b_0 x}{y} \right) = \frac{3}{4\alpha^2} \quad (33)$$

The constant α is determined from the time integral

$$T_{12} = \int_0^{T_{12}} dt = 2\int_{y=0}^{y=-y_1} dy/z = 2\pi\alpha \quad (34)$$

A relation is desired between x_1 and y_1; moreover the function $h(E) = h(x)$ is still

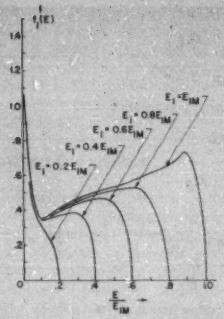


Fig. 14. Nonlinear function $h_1(E)$ for various magnitudes of initial error, fourth-order system

to be determined. The quantities can be found most easily by proceeding directly to the time solution for the nonlinear system

$$\int_0^t dt = \int_{y=0}^{y} dy/z \tag{35}$$

Substitution of equation 31 into equation 35 yields

$$y = y_1/2\left(\cos\frac{2\pi t}{T_{\pi}} - 1\right) \quad (0 \le t \le T_{\pi}) \tag{36}$$

Integration of equation 36 then results in

$$x = x_1\left(1 + \frac{1}{2\pi}\sin\frac{2\pi t}{T_{\pi}} - \frac{t}{T_{\pi}}\right) (0 \le t \le T_{\pi}) \tag{37}$$

Evaluating equations 36 and 37 at $t = T_{\pi}/2$, and $x = x_1/2$ gives $y_1/x_1 = 2/T_{\pi}$.

The synthesis procedure has, thus far, produced the nonlinear differential equation

$$E''' + h(E)E'' + \frac{3\pi^2}{T_{\pi}^2}E' + \frac{4\pi^2}{T_{\pi}^2}E = 0 \tag{38}$$

The function $h(E)$ remains to be determined. This is accomplished by expressing $h(E)$ in terms of the time solutions and then computing its values at specific times in the interval $(0 \le t/T_{\pi} \le 1)$. Likewise, values of $E = x$ can be computed from equation 37 for the corresponding values of time. From these data a plot of $h(E)$ versus E is constructed. Thus

$$h(E) = \left(-E''' - \frac{3\pi^2}{T_{\pi}^2}E' - \frac{4\pi^2}{T_{\pi}^2}E\right)/E'' \tag{39}$$

or

$$h(E) = \frac{\pi}{T_{\pi}}\left[\left(\sin^2\frac{\pi t}{T_{\pi}} + \frac{1}{\pi}\sin\frac{2\pi t}{T_{\pi}} - \frac{2t}{T_{\pi}}\right)\Big/ \sin\frac{2\pi t}{T_{\pi}}\right] \quad (0 \le t \le T_{\pi}) \tag{40}$$

Fig. 10 illustrates the nature of $h_1(E) = T_{\pi}h(E)/2\pi$ for various magnitudes of initial error.

Next, to establish an equivalence between the nonlinear system of equation 38 and the contactor system described by equation 24, the maximum values of the third derivatives of these two systems are equated assuming a step input of position. The expression for the contactor servo is

$$S = \frac{32E_1}{T_{\pi}^3} \tag{41}$$

while the maximum value of the third derivative for the nonlinear system will be

$$|E'''|_{max} = \frac{4\pi^2 E_1}{T_{\pi}^3} \tag{42}$$

Equating equations 41 and 42 yields

$$T_{\pi} = 1.073^{+}T_{c3} \tag{43}$$

The response time of the class-2 third-order nonlinear servo is, therefore, only 7.3 per cent greater than the response time of the optimum contactor servo. Fig. 11 presents a comparison of the time solutions for these two systems.

The stability of the singular point of the nonlinear system of equation 38 can be determined by linearizing this equation about its singular point which is at the origin of its phase space. The application of Routh's[9] criterion of stability indicates that the singular point will be stable if

$$h_1(0) > 0.212 \tag{44}$$

Any fairly good approximation of the function $h_1(E)$, as illustrated in Fig. 10, will produce a $h_1(0)$ well above the value needed for stability.

The existence of limit cycles in this system for initial values of error greater than E_{1max} cannot be directly determined since the extension of Bendixson's theorem to higher order systems has thus far not been accomplished. However, if the system designer adjusts the function $h_1(E)$ to have a constant value B for errors greater than the maximum initial error expected, then provided $B > 0.212$, the system will be described by a stable, linear differential equation in this region of operation.

The coefficients of the synthesized nonlinear differential equation 38 can now be adjusted so that the behavior of the nonlinear system for varying magnitudes and algebraic sign of the initial error will correspond to the behavior of the equivalent contactor system. Making the substitution

$$T_{\pi} = T_{\pi M}(E_1/E_{1M})^{1/3} \tag{45}$$

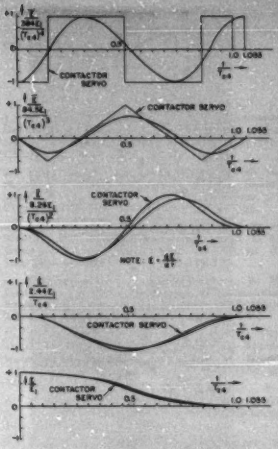


Fig. 15. Time solutions for fourth-order contactor and nonlinear servomechanisms

into equation 38 yields

$$E''' + \frac{2\pi}{T_{\pi M}}\left(\frac{E_{1M}}{E_1}\right)^{1/3} h_1(|E|)E'' + \frac{3\pi^2}{T_{\pi M}^2}\left(\frac{E_{1M}}{E_1}\right)^{2/3} E' + \frac{4\pi^2}{T_{\pi M}^2}\frac{E_{1M}}{E_1}E = 0 \tag{46}$$

The physical servo system represented by equation 46 is shown in Fig. 12. Here, it is found that

$$G_1 = \frac{4\pi^2 JT_A}{T_{\pi M}^2 K_A}\frac{E_{1M}}{E_1} \tag{47}$$

$$G_2 = \frac{3\pi^2 JT_A}{T_{\pi M}^2 K_A}\left(\frac{E_{1M}}{E_1}\right)^{2/3} - \frac{F}{K_A} \tag{48}$$

$$G_3 h(|E|) = \frac{2\pi JT_A}{T_{\pi M}K_A}\left(\frac{E_{1M}}{E}\right)^{1/3} h_1(|E|) - \frac{J}{K_A} - \frac{FT_A}{K_A} \tag{49}$$

Fourth-Order System

Fig. 13 gives the αw, s-plane properties of the phase trajectory for the fourth-order contactor servo along with the approximating trajectory for the nonlinear system. Only the results of the synthesis process will be presented here; a complete derivation is given in the literature.[10]

The synthesized nonlinear differential equation, with its coefficients adjusted for varying magnitude and sign of the initial error and with the maximum value of its fourth derivative equated to the maximum value of the fourth derivative

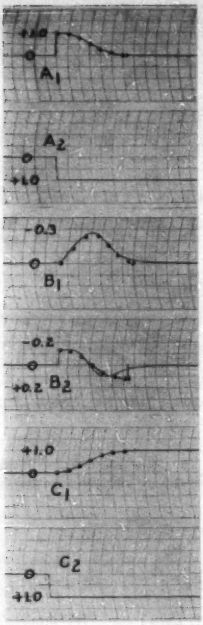


Fig. 16. Recorded response of the synthesized second-order nonlinear servo to a step input of position

A_1—$u(t)$
A_2—$u(0)=1$
B_1—u'
B_2—u''
C_1—$u(0)-u(t)$
C_2—$u(0)=1$
000—Calculated theoretical response
———Computer output
Time scale—5 millimeters = 1 second

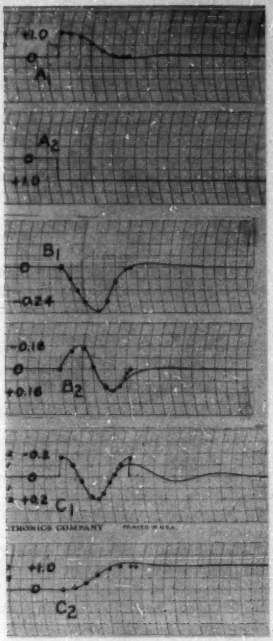


Fig. 17. Recorded response of the synthesized third-order nonlinear servo to a step input of position

A_1—$u(t)$
A_2—$u(0)=1$
B_1—u'
B_2—u''
C_1—u'''
C_2—$u(0)-u(t)$
000—Calculated theoretical response
———Computer output
Time scale—5 millimeters = 1 second

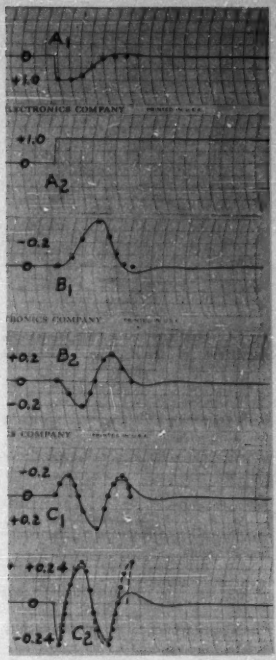


Fig. 18. Recorded response of the synthesized fourth-order nonlinear servo to a step input of position

A_1—$u(t)$
A_2—$u(0)=1$
B_1—u'
B_2—u''
C_1—u'''
C_2—u''''
000—Calculated theoretical response
———Computer output
Time scale—5 millimeters = 1 second

of the optimum contactor servo, is given by

$$E''''+\frac{2\pi}{T_{\alpha M}}\left(\frac{E_{1M}}{E_1}\right)^{1/4}h_5(|E|)E'''+\frac{80}{T_{\alpha M}^2}\left(\frac{E_{1M}}{E_1}\right)^{1/2}E'+\frac{195}{T_{\alpha M}^3}\left(\frac{E_{1M}}{E_1}\right)^{1/4}E'+\frac{476}{T_{\alpha M}^4}\frac{E_{1M}}{E_1}E=0 \quad (50)$$

The system of equation 50 will have a stable singular point provided that

$$0.417<h_5(0)<4.81 \quad (51)$$

A plot of $h_5(E)$ is given in Fig. 14, while the time solutions are shown in Fig. 15. The equivalence relationship is given by

$$T_{4c}=1.055T_{c4} \quad (52)$$

since the maximum value of the fourth derivative for the contactor servo can be expressed as

$$S=384E_4/T_{c4}^4 \quad (53)$$

Computer Study

Figs. 16 through 18 represent results obtained from an analogue computer study of the second-, third-, and fourth-order nonlinear systems synthesized herein. In each case, the function $h(E)$ was approximated by a simple curve and the response time was arbitrarily selected to be 2π seconds leading to the following nonlinear differential equations

$$u''+(1-u)^3u'+0.125u=0 \quad (54)$$

$$u'''+(1-u)u''+0.75u'+0.159u=0 \quad (55)$$

$$u''''+f_5(u)u'''+2.023u''+0.786u'+0.306u=0 \quad (56)$$

$$f_5(u)=0.5+0.5(u-0.5)-6(u-0.5)^3 \quad (57)$$

In spite of the rather inexact approximations for the nonlinear term used in this study, there is good agreement between computer results and the exact results obtained from the time solutions of the synthesized nonlinear differential equations. The amount of discrepancy

between the theoretical and computer results is limited only by the precision of this approximation. Symmetrical response for positive or negative initial values of u is obtained by replacing u with $|u|$ in the nonlinear functions of equations 54, 55, and 57.

Summary Conclusions

The purpose of this paper is to present a process for synthesizing nonlinear differential equations describing the behavior of stable, physically realizable servomechanisms when stimulated by step inputs of position. The phase trajectory of the optimum contactor servo is used as a guide in determining the limiting position of the approximating trajectory for the nonlinear system. Theorems are provided for testing a selected approximating trajectory for synthesis suitability. Time solutions of the synthesized nonlinear differential equations are obtained in a number of cases.

The class-2 nonlinear servo is of particular interest because of the simplicity of its time solutions and because its response time is only a few per cent greater than the response time of the equivalent contactor system. When the nonlinear functions of the class-2 servos are approximated by simple curves (which would apply to physical systems), the resulting nonlinear system exhibits acceptable response characteristics which are in good agreement with the theoretically predicted results.

Although this study is limited to step inputs of position, there is no reason why the philosophy of synthesis cannot be extended to include other types of inputs. Moreover, the authors believe the synthesis approach to nonlinear problems, presented herein, has a much wider field of application.

Appendix I. Proof of Theorem I and Corollary

Equation 4 is written as

$$a_1 = -\frac{y}{x}\frac{dy}{dx} - \frac{y}{x}f(x) \tag{58}$$

Passing to the limit yields

$$a_1 = -\frac{1}{x_1}\lim_{\substack{x\to x_1\\ y\to 0}}\left(y\frac{dy}{dx}\right) \tag{59}$$

Thus the constant a_1 is determined since, by hypothesis

$$\lim_{\substack{x\to x_1\\ y\to 0}}\left(y\frac{dy}{dx}\right)$$

exists and

$$\lim_{\substack{x\to x_1\\ y\to 0}}[yf(x)] = 0$$

Solving equation 58 for $f(x)$ results in

$$f(x) = -\frac{dy}{dx} - a_1\frac{x}{y} = -\frac{1}{\alpha}\frac{d\varphi(x)}{dx} - a_1\frac{\alpha x}{\varphi(x)} \tag{60}$$

Hence, because of the assumed continuity of $\varphi(x)$ and $d\varphi(x)/dx$, $f(x)$ is determined and will be continuous except, perhaps, at points where $y = 0$

$$T_{1,2} = \left|\int_{x_1}^{x_2} dt\right| = \left|\int_{x=x_1}^{x=x_2} dx/y\right| \tag{61}$$

$$T_{1,2} = \alpha\left|\int_{x_1}^{x_2} dx/\varphi(x)\right| = \alpha F_1(x_1, x_2) \tag{62}$$

Hence

$$\alpha = F_1(x_1, x_2, T_{1,2}) \tag{63}$$

but

$$a_1 = -\frac{1}{x_1}\lim_{x\to x_1}\left(\frac{\varphi(x)}{\alpha^2}\frac{d\varphi(x)}{dx}\right) \tag{64}$$

Thus

$$a_1 = g(x_1, x_2, T_{1,2}) \tag{65}$$

Similarly, from equation 60

$$f(x) = -\frac{1}{\alpha}\frac{d\varphi(x)}{dx} - a_1\frac{\alpha x}{\varphi(x)} \tag{66}$$

so that

$$f(x) = h(x, x_1, x_2, T_{1,2}) \tag{67}$$

Appendix II. Proof of Theorem II

Equation 25 is equivalent to the system of equations

$$\frac{du_{p-1}}{dt} + h(u_0)u_{p-1} + b_{p-2}u_{p-2} + \ldots + b_1u_1 + b_0u_0 = 0 \tag{68}$$

$$\frac{du_{i-1}}{dt} = u_i \qquad (i = 1, 2, \ldots, p-1) \tag{69}$$

The combination of equations 68 and 69 for $i = (p-1)$ yields

$$u_{p-1}\frac{du_{p-1}}{du_{p-2}} + h(u_0)u_{p-1} + b_{p-2}u_{p-2} + \ldots + b_1u_1 + b_0u_0 = 0 \tag{70}$$

Now, since by premise 1 the trajectory begins at the point $(u_{0_0}, 0, 0, \ldots, 0)$, the evaluation of equation 70 at this point along with the application of premises 6 and 7 results in

$$b_0 = -\frac{A_0}{u_{0_0}} \tag{71}$$

Substitution of the co-ordinates of the additional $(p-2)$ points at which premises 6 and 7 hold into equation 70 will produce exactly $(p-2)$ independent equations involving the $(p-2)$ unknown constants $b_1, b_2, \ldots, b_{p-2}$. Therefore, these constants are determined from the simultaneous solution of the $(p-2)$ equations.

Solving for $h(u_0)$ in equation 70

$$h(u_0) = \frac{-1}{u_{p-1}}\left(u_{p-1}\frac{du_{p-1}}{du_{p-2}} + b_{p-2}u_{p-2} + \ldots + b_1u_1 + b_0u_0\right) \tag{72}$$

Premises 2, 3, and 4 assure the continuity of $h(u_0)$ except, perhaps, at points where u_{p-1} equals zero.

References

1. Nonlinear Techniques for Improving Servo Performance, Donald McDonald. *Proceedings*, National Electronics Conference, Inc., Chicago, Ill., vol. 6, 1950, pp. 400–42.

2. The Use of Nonlinear Feedback to Improve the Transient Response of a Servomechanism, J. B. Lewis. *AIEE Transactions*, vol. 71, pt. II, 1952 (Jan. 1953 section), pp. 449–53.

3. A Differential-Analyzer Study of Certain Nonlinearly Damped Servomechanisms, R. R. Caldwell, V. C. Rideout. *Ibid.*, vol. 72, pt. II, July 1953, pp. 165–70.

4. Principles of Automatic Control, R. M. Howe. Course notes, Department of Aeronautical Engineering, University of Michigan, Ann Arbor, Mich., pp. (11–31)–(11–35).

5. An Investigation of the Switching Criteria for Higher Order Contactor Servomechanisms, Irving Bogner, Louis F. Kazda. *AIEE Transactions*, vol. 73, pt. II, July 1954, pp. 118–27.

6. Jacobian Elliptic Function Tables (book), L. M. Milne-Thompson. Dover Publications, Inc., New York, N. Y., 1950.

7. Problème Général de la Stabilité du Mouvement (General Problems of Stability of Motion), A. Liapounoff. *Annales de Toulouse*, Paris, France, vol. IX, 1907.

8. Sur les Courbes Définies par des Équations Différentielles (Curves Defined by Differential Equations), I. Bendixson. *Acta Mathematica*, Stockholm, Sweden, vol. 24, 1901.

9. Advanced Part of the Dynamics of a System of Rigid Bodies (book), E. J. Routh. Macmillan & Company, London, England, vol. II, 1930.

10. A Phase Space Method for the Synthesis of Nonlinear Servomechanisms, R. E. Kuba. *Doctoral Dissertation*, University Microfilms, University of Michigan, Ann Arbor, Mich., 1955.

Discussion

I. Flügge-Lotz (Stanford University, Stanford, Calif.): I agree with the authors that the introduction of nonlinearities into servomechanisms may provide a better response. In most cases, like in the one presented here, a specified response to a step input determines the nonlinearities to be used. C. F. Taylor and I recently developed a system[1] the response of which is not based on a specified input, but which adjusts itself to the qualities of any input by providing eight linear systems and by switching from one to another when the deviation from the input advises it results are remarkably good, even for random inputs.

Reference

1. Synthesis of a Nonlinear Control System, I. Flügge-Lotz, C. F. Taylor. *Transactions*, Professional Group on Automatic Control, Institute of Radio Engineers, New York, N. Y., May 1956, pp. 3–9.

Thomas J. Higgins (University of Wisconsin, Madison, Wis.): I read this paper with the greatest of interest. In preparing my paper[1] for the 1956 Princeton Conference, I recently reviewed the literature exhaustively

encompassed in the roughly 500 items of the complementary "Bibliography of Nonlinear Control Systems Theory" (a copy can be obtained free of charge from Director, Engineering Experiment Station, Mechanical Engineering Building, University of Wisconsin, Madison, Wis.). I do not recall reading therein any procedure for synthesizing systems approximating optimum contactor response (as defined by the authors) which parallels that advanced by the authors. Further, I do not recall therein any procedure which conjoins the flexibility, ease of use, and goodness of result entailed in their method. For those who would use this method it may be noted that the theory underlying it is developed most lucidly, whence it can be grasped with ease. Finally, the gratifying agreement between responses obtained by computer and by theoretical calculation both evidences the accuracy of the latter and the proper functioning of the former. I wish to compliment the authors on their successful accomplishment of this fine piece of analytic work in a difficult phase of control theory.

Reference

1. A Résumé of the Development and Literature on Nonlinear Control System Theory, T. J. Higgins. *Transactions*, American Society of Mechanical Engineers, New York, N. Y., vol. 78, 1956.

D. C. McDonald (Cook Research Laboratories, Skokie, Ill.): It has been a pleasure to read this paper. The technique evolved should be extremely useful in synthesizing high-performance servomechanisms which have been intentionally nonlinearized.

I have come across a few points which I would like to have clarified. The authors state that: "The ultimate objective of introducing nonlinearities into a servo system is to produce an optimum response characteristic exemplified by the equivalent contactor servomechanism of the same order." Based on heuristic reasoning it can be concluded that the dynamic response characteristics of a contactor servomechanism appear to be the optimum ones for a large number of different types of input commands. Could the authors supplement such reasonings by some reference to a formalized proof of this optimum response characteristic?

It is also worth while pointing out that there are cases where nonlinearities have been intentionally added to a servo system to improve the steady state as well as the transient response of the servomechanism. For these cases, the nonlinearities introduced were not intended to make the system response approach that of a contactor but rather to obtain a compromise between optimum transient and steady-state characteristics. However, I believe it is true that the great majority of intentionally nonlinearized systems are made nonlinear in an effort to achieve an improvement in the dynamic response characteristics.

The analysis and examples in the paper definitely indicate that the general form of equation 1 will yield a nonlinear system which approximates quite adequately the performance of a contactor servomechanism. However, would the authors explain why the form of equation 1 was chosen, and whether another form might work equally well? Such an additional form would thus offer an alternate approach for the engineer carrying out the synthesis problem. The technique employed in equation 1 of utilizing only one coefficient which is a function of the error does help the synthesis problem considerably. However, must all of the remaining coefficients of the lower order terms be functions of initial conditions, etc.?

The form of the family of trajectories chosen for equations 2, 3, 29, and 30 appears to be adequate from two standpoints:

1. When $n=1$, the trajectories become those of the contactor system.
2. When n is other than 1, the trajectories for these equations satisfactorily represent the contactor-type trajectory.

However, could the basis for choice of the forms of these four equations be explained? Also, could other forms be utilized which would thus afford the design engineer an alternate approach?

I would like to repeat again that I believe that the synthesis technique outlined in this paper is extremely valuable and appears to be a proper approach for determining how to nonlinearize intentionally a servomechanism.

R. E. Kuba and **L. F. Kazda**: We are glad to have our attention called to reference 1 of Mrs. Flügge-Lotz's discussion. It is our recollection that the system described therein was a general nonlinear second-order system describable by the following nonhomogeneous equation

$$y''+b(y,t)y'+c(y,t)y=f(t)$$

The functional dependence of the coefficients was so chosen as to minimize the instantaneous error for an arbitrary forcing function $f(t)$. The objectives that we had in mind in our approach were:

1. To develop a system which would require no switching knowledge.
2. To develop a method of synthesis which would be applicable to higher order systems.

Although this study is limited to step inputs of position, the procedure of synthesis presented herein can be extended to include other types of inputs. An experimental model constructed employing the results embodied in this paper has demonstrated that substantial improvement in performance can be obtained over the performance of an equivalent linear system.

We most gratefully acknowledged Prof. Higgins' remarks. In response to questions raised by Mr. McDonald, it may be said that the literature[1,2] does contain specific formalized proofs for the optimized characteristics attributed to the type of contactor servomechanisms considered in this paper.

The form of equation 1 was chosen because:

1. Sufficient conditions could be established to insure success of synthesis, i.e., unique determination of the coefficients b_0, b_1, $b_2 \ldots b_{p-2}$, and of the function $h(E)$.
2. The differential equation automatically exhibits a stable linear region of response about the origin of its phase space. This is because the function $h(E)$ always approaches positive infinity as the system error approaches zero, regardless of the order of the system. Thus, the function $h(E)$ may always be approximated by a real, positive constant in this region. Therefore, the stability of the singular point is assured.
3. Symmetrical system operation for positive and negative errors is easily obtained by making the function $h(E)$ an even function of the system error.
4. The effects of viscous output damping are easily compensated by proper adjustment of the coefficient b_1 in equation 1 or, in the case of a second-order system, by a translation of the function $h(E)$.
5. As the order of the system increases from p to $(p+1)$, the complexity of the differential equation increases by only one term.
6. Computer results have shown that relatively large changes in the coefficients b_0, $b_1, \ldots b_{p-2}$, and in the function $h(E)$ produce only relatively small changes in system response.

We do believe that other forms of differential equations can be chosen for synthesis. The problem is to determine sufficient conditions which will test a chosen phase trajectory for synthesis suitability. For example, we have considered the nonlinear differential equation

$$\frac{d^nE}{dt^n}+C_{n-1}\frac{d^{n-1}E}{dt^{n-1}}+\ldots+C_1E'+f(E)E'+C_0E=0 \quad (n\geq 2)$$

It is possible to synthesize this equation to fit certain phase trajectories; however the function $f(E)$ will, in general, not be unique, and stable singular points are not assured for odd values of n. The coefficients b_0, b_1, $\ldots b_{p-2}$ will, in general, be functions of the initial conditions, but this should not prove a serious problem to servo designers.

The choice of a phase trajectory used as a starting point for synthesizing a nonlinear differential equation should be guided by these qualifications:

1. Does the selected phase trajectory pursue a desired path in the phase space?
2. Can the topology of the phase trajectory be easily adapted for extension to higher order systems?
3. Will the selected phase trajectory pass the test for synthesis, e.g., theorems I and II?

These considerations have been the basis for the choice of equations 2, 3, 29, and 30. Other forms could be utilized, provided a test for synthesis is known and the test is passed. Thus, any trajectory satisfying the conditions imposed by theorems I and II would produce satisfactory synthesis of equation 1.

References

1. Differential Equations with a Discontinuous Forcing Term, D. W. Bushaw. *Ph.D. Thesis*, Princeton University, Princeton, N. J., 1952.
2. See reference 5 of the paper.

Applying Industrial Controllers for Safe Operation

G. W. HEUMANN
MEMBER AIEE

THE DESIGNER of controllers and control equipments has three major design objectives. In satisfying these objectives, he discharges his obligations to various sectors of his professional responsibility. These objectives are

1. To design for quality of performance. This objective satisfies the obligation to the customer who expects to receive fair value for the price he pays.

2. To design for economy of manufacture. This objective satisfies the obligation to the manufacturer who must sell at a profit to stay in business.

3. To design for safe operation. This objective satisfies the obligation to the public since it safeguards operators, as well as the owner and the neighbors of the plant, against lethal accidents and fire hazards.

The last objective will be discussed in this paper.

Safe operation is the objective of the "Standard for Industrial Control Equipment"[1] which has been issued by the Underwriters' Laboratories, Inc. (U/L). This standard sets forth the rules under which control equipment is tested for approval and listing under the U/L re-examination service. Controllers designed to meet U/L standards can be presumed to meet their rating and operating overloads without undue heating. Thus, they do not constitute a fire hazard. Clearances and creepages, dielectric tests, and strength of enclosures are so specified that accidental shocking of operators, by inadvertent energizing of supposedly dead parts, through deformation or dielectric breakdown, is avoided.

"Standards for Industrial Control"[2] of the National Electrical Manufacturers' Association (NEMA), in addition to standard ratings, performance standards, and manufacturing standards, contain the same safety provisions as the U/L standard.

The "National Electrical Code"[3] (NEC), although it is primarily an installation code and most of its provisions deal with the manner in which electric equipment is installed, has a bearing on controller design by establishing rules under which the motor branch circuit is protected against operating overloads and short circuits and thus design and selection of overload relays and branch-circuit interrupters is affected. NEC also governs enclosures for controllers operating in certain specified hazardous atmospheres, to safeguard against explosions due to electric arcs.

Controllers as listed in manufacturers' catalogues and designed to meet U/L and NEMA standards, as well as NEC, can be deemed to be inherently safe when properly selected to meet operating conditions. However, a controller which is perfectly safe when installed in one plant may become a lethal hazard when installed in some other plant. It is the responsibility of the application engineer to select, from a number of standard or special designs, the one which will obtain the highest degree of protection under the specific plant conditions.

From the point of view of safety the most important considerations are: selection of enclosures to avoid fire hazard due to contamination of industrial atmosphere; and co-ordination of controller interrupting ability with the short-circuit capacity characteristic of the power system, to avoid the possibility of explosive failure of equipment that might endanger the operator and surrounding equipment, and create a fire hazard.

Enclosures for Industrial Use

Open-type controllers should be used only in clean places where they are installed in a room separated from the operating area, and accessible only to competent service personnel. All controllers to which operating personnel have access should be enclosed and have lockable doors to prevent possible accidental contact with live parts and consequent hazard to life and limbs of inexperienced personnel.

General-purpose NEMA type *1* enclosures are perfectly satisfactory when the controller is installed in a clean environment. When the atmosphere contains flammable contaminations of such concentration that an electric arc may touch off an explosion, NEC requires the use of special enclosures (NEMA types *7*, *8*, *9*, *9A*) the design of which is governed by U/L standards. Such enclosures are comparatively large and expensive, and it would be uneconomical to use them in contaminated, but not flammable, atmospheres.

In between clean and flammable atmospheres there are those which contain dust, lint, fibers and flyings, or oil and coolant seepage. The hazards which these contaminations entail are as follows:

1. Dust may contain conducting particles which, when accumulated inside the enclosure, cut down electrical clearances. Thereby an operator may be shocked, or a short circuit may become a fire hazard.

2. Nonconducting combustible materials such as dust, lint, fibers, flyings (as may be found, for instance, in textile and tobacco plants), may accumulate around the outside of the controller enclosure and may also seep inside. With an enclosure which is not tight, accumulations within the case may be ignited by electric arcs, and the resultant flame communicated to the outside accumulations through cracks between case and cover or unused conduit and mounting holes. A fire may thus be started.

3. Seepage of oil or coolant into the enclosure may deteriorate insulating materials and cut down clearances. As a result, the operator may be shocked, or a short circuit may initiate a fire hazard.

To eliminate these hazards NEMA has introduced the "Type *12* Industrial Use Enclosure" which is defined in its standards as follows:

"A type *12* enclosure is designed for use in those industries where it is desired to exclude such materials as dust, lint, fibers and flyings, or oil and coolant seepage."

Type *12* enclosures have an oil-resistant synthetic gasket between the case and the cover, and cover fastener parts are captive so that they do not become lost when the cover is opened. There are no conduit openings. Larger enclosures for combination starters and multiple controllers consist of a case which is completely tight in the back with no holes at all for mounting the controller in the enclosure. External mounting feet are provided. Fig. 1 shows a typical combination starter in type *12* enclosure.

Enclosures of this kind have been used in large number by the automotive and other industries using high-production machinery, with excellent operating experience. Type *12* enclosures are made

Paper 56-269, recommended by the AIEE Industrial Control Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Winter General Meeting, New York, N. Y., January 30-February 3, 1956. Manuscript submitted November 28, 1955; made available for printing May 4, 1956.

G. W. HEUMANN is with the General Electric Company, Schenectady, N. Y.

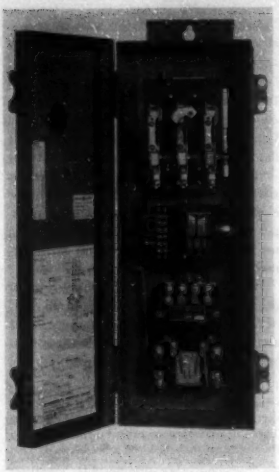
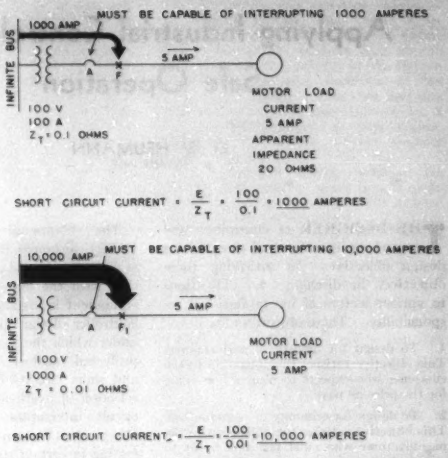


Fig. 1. Combination starter mounted in type 12 enclosure

Fig. 2. Manual starter (loom switch) in type 12 enclosure

Fig. 3. Comparison between load current and short-circuit current



mandatory in the "JIC Electrical Standards for Industrial Equipment."[4] Their use is justified in the "Safety Interpretations" as follows:

"The accumulation of dirt, dust, chips, oil, coolant, or other foreign material in an electrical enclosure causes short circuits or grounds. This results in injury to employees, due to malfunction of the equipment or from explosions and fire. It is the intent of these paragraphs (calling for type *12* enclosure) to eliminate such conditions."

Many smaller controllers, such as small magnetic or manual starters, are used in the textile and other industries which have to contend with fibers and flyings, but not with oil or coolant seepage. A typical application is the loom switch of Fig. 2. For such applications type *12* construction permits mounting holes in the case, but no holes may remain unused.

Type *12* enclosures are being so favorably received in the textile industry that the Subcommittee on Textile Industry of the AIEE Committee on General Industry Applications has prepared "Proposed Recommended Practices for Electric Installations on Textile Machinery," specifying type *12* enclosures for motor controllers.

There are numerous installations in industrial plants which are not covered by specific industrial standards and on which the application engineer has to use his best judgment. If there is any doubt at all concerning the adequacy of general-purpose type *1* enclosures, type *12* enclosures should be given preference.

Adequate Interrupting Capacity Controllers

Another unusual service condition is exposure of control equipment to faults in the motor branch circuit. Thermal capacity to withstand normal load and operating overloads are determined by the controller horsepower rating. Standard general-purpose controllers are designed to interrupt operating overloads up to stalled rotor current. Under NEMA standards general-purpose a-c controllers have an interrupting ability of ten times rated motor current; d-c controllers for full-voltage starting are also capable of interrupting ten times rated motor current, whereas d-c controllers for reduced-voltage starting can interrupt four times rated motor current.

Short-circuit currents which may occur in a motor branch circuit bear no relationship to the rating of the motor which is served by the circuit. The short-circuit current available at the point of fault of a motor branch circuit is largely determined by the size of the transformer feeding the motor branch circuit, as is illus-

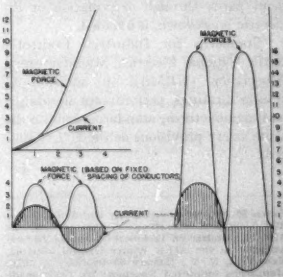
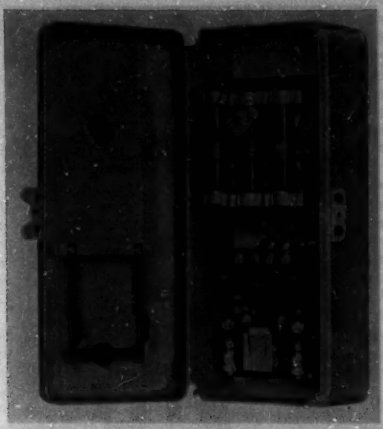


Fig. 4. Magnetic forces due to short-circuit current

Fig. 5. Combination starter with molded-case circuit breaker

Fig. 6. Combination starter with fused motor-circuit switch



trated in Fig. 3. Two cases are shown. In the upper, a motor drawing 5 amperes is connected to the power system through a transformer of small capacity which permits a fault current of 1,000 amperes to flow. In the lower, the same motor is connected to a transformer having ten times the capacity of the former one, hence one tenth its impedance, thus permitting a fault current of 10,000 amperes to flow. This comparison shows that fault currents may flow through a controller, which are greatly in excess of the aforestated interrupting abilities of general-purpose controllers. They would blow up and cause accidents and fires unless steps are taken to design adequate interrupting ability into the combination of motor controller and branch-circuit protector.

NEC requires that each motor branch circuit be equipped with means to protect the branch circuit against fault currents. This can be accomplished by:

1. Building high interrupting ability into the controller itself so that the same set of contacts interrupt both motor load and fault currents. For economic reasons this is not done generally for low-voltage (600-volt and less) controllers, but it is often done on high-voltage (601- to 4,800-volt) controllers.

2. Using a combination of a general-purpose controller, the contacts of which interrupt motor load currents, and a circuit breaker or fuses, which interrupt fault currents. This is the usual practice for low-voltage installations, and it is frequently applied to high-voltage equipments.

Many users believe that a motor branch circuit is adequately protected when a controller with limited interrupting ability is backed up by a circuit breaker having an interrupting rating commensurate with the short-circuit current available from the power system. Such a combination may or may not be safe. To have a safe installation it is necessary carefully to co-ordinate the interrupting characteristic of the circuit interrupter and the thermal and mechanical capacity of the controller. Any circuit interrupter requires a finite time interval between initiation and interruption of the short circuit. During this interval the let-through current flows through the controller, and its i^2t determines the thermal stress upon the conducting parts. If the let-through energy is too great, controller parts may melt and initiate an internal short circuit, which in turn may cause the controller to explode. Of equal importance is the effect of magnetic forces on conducting parts of the controller. These forces vary as the square of the current, as illustrated in Fig. 4. If permitted to persist, they may cause breakage of parts and consequent destruction of the controller.

Do Accidents Occur?

No matter how carefully a power system has been designed and installed, short circuits cannot be avoided entirely. Therefore, the only remedy available is to design the power system so as to mitigate the destructive effects of short circuits when they do occur by using a combination of motor-starter and motor-branch-circuit protection which is capable of successfully interrupting the short circuit without damage to itself.

The fact that no short circuits have occurred so far on a branch circuit is no assurance at all that no short circuit will occur in the future. They can happen at any time and without warning. Therefore, the investment in equipment with adequate short-circuit capacity should be

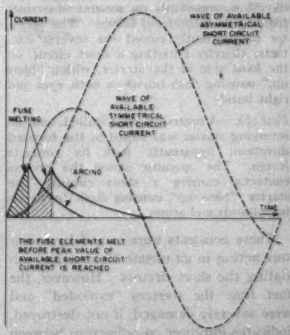

CURRENT
WAVE OF AVAILABLE ASYMMETRICAL SHORT CIRCUIT CURRENT
WAVE OF AVAILABLE SYMMETRICAL SHORT CIRCUIT CURRENT
FUSE MELTING
ARCING
TIME
THE FUSE ELEMENTS MELT BEFORE PEAK VALUE OF AVAILABLE SHORT CIRCUIT CURRENT IS REACHED

Fig. 7. Current-limiting action of high-voltage fuses

Fig. 8. High-voltage high-interrupting-capacity class E2 air break controller

looked upon as buying insurance against potential trouble.

Too few people are conscious of the fact that serious accidents may be caused by failure of low-voltage controllers under short circuit. Unfortunately it is difficult to find reliable and detailed reports on electrical accidents. The following cases are taken from a 1954 report on electrical work injuries in California industries:[5]

Case (a). An operator attempted to reset an overload relay on a reversing starter, the contacts of which were closed in the forward direction, apparently on account of contact freezing or mechanical failure. The operator inadvertently closed the reverse contacts, thereby initiating a short circuit on the load side of the starter, which "blew up," causing flash burns on both eyes and right hand.

Case (b). A reversing drive stalled, and the reversing starter was closed in the forward direction, apparently with its contacts frozen. The operator closed the reverse contacts, causing a short circuit. The starter "blew up" causing flash burns on both hands and arms.

These accidents were caused by operators acting in an unsafe manner and initiating the short circuits. However, the fact that the starters "exploded" and were severely damaged, if not destroyed, indicates improper co-ordination between controller and short-circuit interrupting means.

Far more serious may be accidents due to short circuits on high-voltage systems. Too many plants still contain controllers of inadequate interrupting ability which are backed up by oil-type feeder circuit breakers having a long interrupting time of the order of eight cycles. The total time during which short-circuit current flows may be even longer, depending on relay setting. The following case histories are indicative of the dangers which are present when controllers with inadequate interrupting ability are installed in high-voltage motor branch circuits.

Case (c). A semiautomatic starter with low interrupting ability was used with a 200-horsepower 2,300-volt synchronous motor. An internal fault developed within the motor. When the starter was closed a short circuit was initiated, which was eventually cleared by the feeder breaker. However, the starter was completely burned up before the fault was interrupted. Fortunately, the starter was of dead-front construction which prevented serious injury to the operator.

Case (d). An induction motor was controlled by a 2,300-volt manual oil-immersed autotransformer starter having an interrupting ability of ten times rated motor current. A fault developed somewhere in the branch circuit. When the operator closed the starter, he closed on a short circuit. The starter burst, the oil caught fire, and the operator was killed. The feeder breaker ahead of the starter proved to be of inadequate interrupting capacity, and it exploded too. The fault was finally cleared by a substation breaker on the high side of the transformer.

Because of the potentially more lethal effect of failures of high-voltage controllers, NEMA no longer recognized as standard high-voltage a-c controllers with an interrupting ability of ten times motor current. Very few power systems are in existence today which produce sufficiently low available short-circuit current to make a "ten times" controller plus back-up breaker a safe combination. The only high-voltage controllers which NEMA recognizes as standard are class *E* controllers which have a stated high-interrupting-capacity rating.

High-Interrupting-Capacity Controllers

General-purpose low-voltage starters having an interrupting ability of ten times rated motor current are a standard product, built in quantity. Experience has proved that inverse-time circuit breakers or fuses (as recommended by NEC), when properly co-ordinated with standard starters, obtain satisfactory circuit protection. Circuit breakers or fuses may be installed separately from the starter. However, industrial control manufacturers offer combination controllers which, in addition to the regular control functions, also perform the function of fault protection. In a combination controller standard starters and breakers or fuses are so co-ordinated that the let-through energy while short-circuit current flows is not large enough to cause substantial damage to the controller. The manufacturer attaches an interrupting rating to the combination, expressed as total available short-circuit current at the point of controller installation. Thus the manufacturer assumes the responsibility for proper design and selection of components, and the user has the assurance that the combination of controller and short-circuit protective means adequately protects his motor branch circuit.

Shown in Fig. 5 is a combination starter consisting of a molded-case air circuit breaker and a full-voltage magnetic starter with thermal overload protection, assembled in a common enclosure. The circuit breaker has combination thermal-magnetic trips, providing inverse-time tripping on low overcurrents, and instantaneous tripping on high fault currents. Combination starters up to size 4 have interrupting ratings up to 15,000 amperes, and larger sizes can be obtained with interrupting ratings up to 25,000 amperes. The breaker also serves as a disconnecting means to isolate the starter from the line, for servicing.

Combination starters are also available with motor-circuit switches (which are rated in horsepower and are capable of interrupting stalled-rotor current) as disconnecting means and fuses for fault protection. Such a starter is illustrated in Fig. 6. Many starters are in service which use standard NEC fuses. Such fuses do not have an a-c interrupting rating. Although they are U/L approved under accepted fuse standards, their manufacturers do not guarantee a definite a-c interrupting ability. Operating experience, however, indicates that NEC fuses perform satisfactorily at moderate system capacities, and control manufacturers offer combination starters with NEC fuses having an interrupting rating of 25,000 amperes. In recent years several makes of fuses have been placed on the market for which their manufacturers have established a-c interrupting ratings. Control manufacturers offer combination controllers with an interrupting rating of 50,000 amperes when equipped with high-interrupting-capacity fuses.

High-voltage controllers are rated in terms of available symmetrical short-circuit kilovolt-amperes. They are designated by NEMA as class *E* controllers. Two basic designs are offered. Class *E1* controllers make use of a contactor which has an inherent high-interrupting ability. The contactor performs the dual function of interrupting motor load and overload currents, as well as interrupting fault currents. Contactors may be air break type, or they may be oil-immersed. The complete controller is housed in an enclosure to afford protection to the operator against accidental contact with live parts. Busses and other components in the main power circuit are so designed that they are not damaged by the let-through energy and the resultant magnetic forces. Standard interrupting ratings are 25,000 and 50,000 kva.

Another recognized type is class *E2* which uses a contactor with an interrupting ability of ten times rated motor current for interrupting motor load and overload currents, and fuses for interrupting short-circuit current. Most class *E2* controllers installed today utilize fast-acting so-called current-limiting fuses. The current-limiting action is illustrated in Fig. 7.

The fuse, containing silver fusible wires or ribbons, embedded in quartz sand, melts so fast that current is prevented from rising to its initial peak value during the first half-cycle after the incidence of the short circuit. In Fig. 7 the shaded area indicates current flow prior to melting of the fuse, current then tapering down during the arcing period, actual arc extinction taking place within approximately one half-cycle. Class *E2* controllers can be connected to systems capable of producing a symmetrical available short-circuit kilovolt-amperes of 150,000 kva at 2,200–2,400 volts, or 250,000 kva at 4,000–4,800 volts.

Shown in Fig. 8 is a typical class *E2* controller designed for full-voltage starting of a synchronous motor. The top compartment contains a gang-operated disconnect switch and the current-limiting fuses. The middle compartment houses a swinging panel containing low-voltage field application and auxiliary control devices, as well as the overload relays. The bottom compartment contains the high-voltage air break contactor. Compartments are separated by fixed steel barriers, so that a maintenance man can safely work on the low-voltage compartment without danger of accidental contact with live parts in the high-voltage compartments. A mechanical interference-type interlock prevents opening of the high-voltage compartment doors unless the line contactor is dropped out and power is removed from the motor branch circuit. The operating handle for the disconnect switch also actuates the latch of the disconnect compartment door. Hence access to the fuses for renewal can only be gained with power off, and with the controller isolated from the line, thereby providing a high degree of safety to the maintenance man who may have to service the controller.

The growth of modern power systems which may be expected to continue in the future has enhanced the potential danger of short circuits in motor branch circuits. Careful consideration should be given to the problem of providing equipment with adequate interrupting capacity. While low-voltage controllers may be adequately protected by separate circuit breakers or fuses, the use of combination controllers assures the user that controller and short-circuit interrupting means have been properly co-ordinated to realize a stated interrupting rating. On high-voltage motor circuits the former practice of using a controller with only ten times rated motor current interrupting ability with an 8-cycle back-up breaker is to be discouraged, as being wholly inadequate for safe operation. Nothing but class *E* controllers with high interrupting capacity should be considered for controlling high-voltage motors.

References

1. Standard for Industrial Control Equipment. Underwriters Laboratories, Inc., Chicago, Ill., Subject 508, seventh edition, 1938.

2. Standard for Industrial Control. *Standard No. IC-1*, National Electrical Manufacturers Association, New York, N. Y., 1954.

3. National Electrical Code. National Fire Protection Association, Boston, Mass., 1956.

4. JIC Electrical Standards for Industrial Equipment. *Electrical Manufacturing*, New York, N. Y., May 1954, pp. 111-28.

5. Electrical Work Injuries in California Industries—Year Ended December 31, 1954. Department of Industrial Relations of the State of California.

Electric System for a Nuclear Reactor

E. P. PEABODY
MEMBER AIEE

D. S. BRERETON
MEMBER AIEE

THE ELECTRIC system for supplying power to a nuclear reactor and its auxiliary equipment in most respects is very similar to that found in the average industrial plant. The main factor that affects the design of such a system is the high degree of reliability necessary to assure reactor safety during all phases of operation, including shut-down periods. The system presented here was recently constructed at the Atomic Energy Commission's Hanford Works and describes one of two identical plants constructed at the site. The General Electric Company is the prime contractor responsible for the engineering and operation of all production facilities at Hanford. C. T. Main, Inc., Boston, Mass., served as engineering contractor on the water plant and its electric system.

The major power requirements of a nuclear reactor facility of the Hanford type are for pumping cooling water. Large quantities of water must pass through the reactor to carry away the intense thermal energy generated by the nuclear reaction. Should this flow of water stop, the internal temperatures could reach destructive values in a matter of seconds. Even when a reactor is shut down, a large amount of heat is released by radioactive decay of fission products in the reactor, thus necessitating substantial amounts of cooling water at all times.

The transition period during shutdown of a reactor is very critical and requires special consideration. When a signal is sent to the reactor calling for a shutdown, an appreciable length of time is required to activate the safety devices and bring about reduction of the power level of the reactor. During this period, normal flow of water is required to prevent overheating of the reactor. Should the cause for shutdown be total loss of pumping energy, an ordinary pumping unit would quit pumping water before the power level of the reactor was reduced and serious damage to the reactor would result. Some means must be provided therefore, to maintain an adequate flow of water during this transition period.

Paper 56-674, recommended by the AIEE Chemical Industry Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Summer and Pacific General Meeting, San Francisco, Calif., June 25-29, 1956. Manuscript submitted March 26, 1956; made available for printing June 25, 1956.

E. P. PEABODY is with the General Electric Company, San Jose, Calif., and D. S. BRERETON is with the General Electric Company, Schenectady, N. Y.

Basic Requirements

The preceding discussion resolves into four basic requirements which must be met if a power system for a nuclear reactor plant is to be adequate. These are:

1. An adequate source of power for normal water pumping and auxiliary loads.
2. The most reliable source of power within practical limitations for supplying water following a shutdown.
3. Means of providing adequate water flow during the transition period of a shutdown.
4. Safety devices which are highly sensitive, accurate, and of the fail-safe type, which will shut down the reactor in an emergency.

Only the first three items will be considered herein. The fourth item is not pertinent to the design of the main electric system and has been covered in another article.[1]

In the following discussion the power source for the major production pumping load will be referred to as the normal system. The power source for supplying the emergency pumping and other critical loads required during shutdown will be referred to as the emergency system.

The normal system must have a high degree of reliability, but interruption of power under extreme conditions will be tolerated. The emergency system must be of the highest order of reliability and must at all times be capable of handling all critical and emergency loads. The emergency system is to be designed so that no more than one unit of load will be lost because of failure of any one system component such as a motor, circuit breaker, bus, feeder, transformer, or generator. Provision against the failure of two separate components of the system is not to be considered when such failure is due to different or unrelated causes.

Description of Water Plant

A description of the hydraulic portion of the plant will help to clarify the requirements imposed on the electric system. Fig. 1 shows a typical flow diagram for normal water flow with one unit of each major component represented. The raw water pump raises water from the Columbia River to a head house where it receives chemical treatment, and then passes through settling basins and filters to a clearwell. By means of primary and secondary process pumps, the filtered water is pumped through the reactor and then discharged to retention basins where it is held until the radioactivity has decayed to a safe value; the water then can be returned to the river.

Power to the raw water pumps can be interrupted for short periods of time without affecting plant operation because of storage capacity in the filter plant and clearwell. The primary pumps provide the necessary positive suction head for the secondary pumps which, in turn, develop the required pressure to deliver water to the reactor. Continuous operation of both pumps is necessary for reactor operation. Power must be available to the primary pumps at all times, however, since these pumps also provide water flow for shutdown periods. Fig. 1 shows flow during both operating and shutdown periods. Note that the primary pump flow by-passes the secondary pump when the latter is not operating. This is automatic through a check valve which opens following pressure decay of the secondary pump.

To provide the necessary transition flow during shutdown following loss of normal power or to permit riding through short transient faults which clear by high-speed relaying, flywheels were provided on the secondary pump shafts. A flywheel giving the system a moment of inertia about six times that of a conventional unit was required to co-ordinate properly the water flow with the shutdown transition cycle. Flywheels were also provided on the primary pumps to maintain an adequate positive suction head for the secondary pumps during extended system faults.

From the foregoing it is seen that the normal system will supply power to the raw water pumps and secondary process pumps while the emergency system must be capable of handling the primary pumps and all critical control, lighting, and auxiliary loads. To provide power for replenishing the clearwell and filters in the event of an extended outage on the normal system, provision is made for at least one raw water pump to be supplied from the emergency system.

Power Sources

The natural choice for the normal system supply was to take power from an existing 230-kv transmission system in

the Hanford area. This system receives power directly from the Bonneville Power Administration's (BPA) Midway substation which connects directly through BPA lines to Coulee and Bonneville Dams. Power costs from this source are less than for power that might be generated locally. The reliability of service has been of a very high order (one total outage in 10 years), though power failure can occur during extreme system disturbances such as instability caused by relay or circuit-breaker misoperation or the simultaneous loss of several lines. Power from this source was felt to be fully adequate for the normal supply, though it did not possess the reliability required for the emergency supply. The existing 230-kv lines had sufficient reserve capacity to carry the additional load.

Since the 230-kv system was the only existing power source in the area it was necessary to supply energy for the emergency functions from within the new plant. Alternates considered were:

1. The use of steam turbine drives on all emergency pumping units with a small steam-driven generator for supplying emergency auxiliary electric loads; steam would be supplied from a central boiler plant.

2. The use of diesel drives on all emergency pumping loads with a diesel-driven generator for emergency auxiliary electric loads.

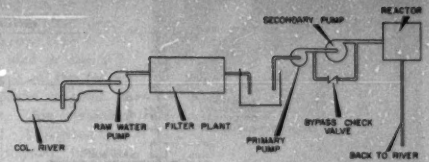
3. The provision of electric drives on all emergency pumping units with local steam-turbine-, diesel-, or gas-turbine-driven generator units to meet their electric power needs.

After careful study, the choice was made to use electric drives on all pumping units and to provide three steam turbine generators to supply emergency power needs, each with its own oil-fired boiler.

Electric System Layout

Fig. 2 is a 1-line diagram of the system. The 230-kv portion connects in series with an existing 230-kv loop which is supplied from BPA's Midway substation. Approximately 4 miles of steel tower line was constructed for this connection. A voltage of 13.8 kv was selected for the secondary of the main power transformers. This voltage is used in other plants in the area and permits the 230–13.8-kv power transformers to be interchangeable. It is also a practical voltage to use on motors of 10,000-horsepower (hp) rating. The remainder of the system operates at 4.16 kv. This permits utilizing 5-kv-class switchgear to its full rating. It is a desirable voltage to use for the generators selected and can be employed for supplying all motors of

Fig. 1. Water flow diagram



250 hp and larger. All minor loads for power and control are supplied at 480 volts with 120/208 volts for lighting.

Method of Operation

Methods of operation were given careful consideration in planning the system. The main points concerned whether or not the local or emergency generators should operate in parallel with the normal system and whether the emergency loads should be carried normally on the emergency system or on the normal system. The final recommendation was for the generators to operate in parallel with the normal system. A recommended method of arranging loads is given in Fig. 2, where the circuit breakers shown in heavy outline are normally closed. In the interest of economy, the generators normally carry a kilowatt load of 10 to 20 per cent of their rating. They are overexcited to carry about 50 per cent of their rating in reactive power during parallel operation. Two generators are normally operating and are capable of carrying all emergency loads. The third generator serves as a spare. The transformer banks connecting the 4.16-kv and 13.8-kv systems also have adequate capacity to supply all loads on the 4.16-kv system. Those loads supplied from busses *B*, *C*, *E*, and *F* comprise the emergency portion of the system, and all loads supplied from busses *A* and *D* and the two 13.8-kv busses are supplied only from the normal system and are subject to interruption.

An understanding of the detailed requirements for emergency power will help to show why the foregoing method of operation is preferred. The primary pumps provide the necessary suction head for the secondary pumps during normal operation. Power to the primary pumps cannot be interrupted for more than about 1 second without danger of cavitation in the secondary pumps. This requires then that the emergency source of power be immediately available and that fast switching be provided if the loads are not normally connected to the emergency source. To meet this requirement means that the generators must be spinning at all times since several seconds are required for them to accelerate and pick up load. The next determination to make is whether the generators should operate in parallel with the system or separate. The choice of separate or parallel operation with the normal system was considered next.

Referring again to Fig. 2, consider that the circuit breakers shown with heavy outline are closed for the condition of parallel operation. For separate operation, assume that busses *A*, *B*, *C*, and *D* are supplied from the normal system by transformers through circuit breakers *16*, *32*, *41*, and *17*, and busses *E* and *F* are supplied by generators 1 and 3 through breakers *65* and *70*. Assume also that bus tie breakers *36*, *37*, *67*, and *68* are closed. For the latter case, the loads may connect to either source of power depending on the amount of load to be carried by the generators.

Separate operation of the two systems requires that critical loads be transferred from one source to the other in the event the first source fails. Assume for one case that three primary pumps are connected to each source and the breakers *30*, *42*, *44*, *60*, *62*, and *74* are closed. Assume also that the other loads are distributed in a similar manner between the two sources. If the normal supply should fail, it would be necessary for breakers *30*, *42*, and *44* to open and breakers *61*, *73*, and *75* to close to assure continued operation of the emergency loads. Failure of any one circuit breaker to function would mean loss of that particular unit of load. Besides the possibility of breakers failing to operate properly, certain hazards are present when effecting rapid transfer of induction motors from one voltage source to another. If the transfer is too rapid, residual flux in the motor rotor can cause severe transient torques which could shear the motor shaft. With the case at hand, most of the motors have about a 2-per-cent slip at full load which results in a shorter decay time for the residual flux than with normal slip motors. With a transfer time of 1/2 to 1 second, this factor does not appear to be a serious problem but, nevertheless, should be avoided since it could cause abnormal transient voltages and stresses in the motors.

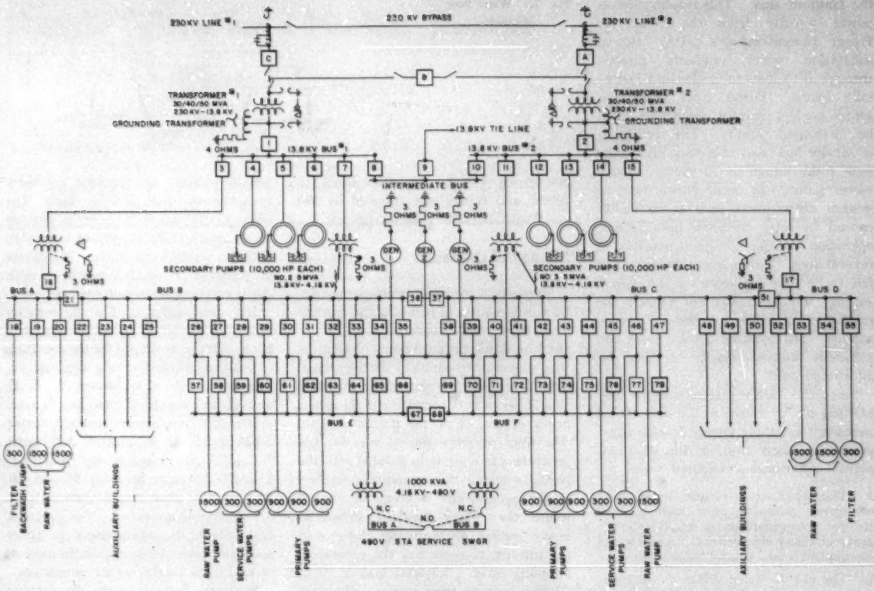


Fig. 2. One-line diagram of reactor plant electric system

The sudden transfer of emergency loads to the generator could cause quite a transient shock to the generators, depending upon the amount of load transferred and the loading of the generators. With the generators lightly loaded, a relatively weak field would be used to limit terminal voltage and a large load suddenly applied would cause considerable voltage drop before the voltage regulators could act. This could cause a reduction in water flow and possibly affect the stability of the emergency system.

With the generators operating separately from the normal system and carrying all emergency loads, no transfer or switching would be required because of disturbances on the normal system. It would be necessary, however, to provide transfer from the emergency to the normal system in the event of trouble on any portion of the emergency system. This would carry the same complications and hazards as pointed out in the foregoing. This method of operation is also less desirable because of the higher cost of steam-generated electricity.

In considering parallel operation of the generators and the normal system, many problems associated with separate operation disappear while others become more predominant. With parallel operation it is only necessary to open four circuit breakers (*7*, *11*, *32*, and *41*) to separate the normal and emergency systems in the event of prolonged trouble on the normal system. Two breakers open in series for each tie point so that one breaker can fail to open without jeopardizing the system. This means that, for a typical separation, only four breakers would be required to open whereas, with separate operation, several breakers must open and an equal number must close. With parallel operation there will be no transient voltage or torque disturbances to the motors or generators during switching.

Since the generators will be in synchronism with the normal system, they can satisfactorily carry any percentage of load desired up to their rating. Parallel operation also permits the use of the generators to control system peaks. Since power billing is on the basis of annual maximum demand, this is an important factor in controlling power costs. The generators can also be used to control voltage or power factor through regulation of their fields. Less voltage drop will be encountered when starting motors since the capacity of both the generators and the normal system will be available to supply starting inrush current.

In summarizing the two methods of operation, the following advantages and disadvantages are listed.

The advantages of parallel operation are:

1. No momentary loss of power to any essential load results when the normal power source is disconnected.
2. No significant transient voltage or torque disturbances to motors or generators.
3. A relatively simple and reliable control and relaying system can be used.
4. Reliability is increased for the most prevalent types of faults since no circuit breakers are required to close and fewer breakers are opened.
5. Control of system load peaks, power factor, and voltage can be achieved.
6. Less voltage drop results when motors are started since more capacity is available to the busses.
7. Switching errors are less apt to cause loss of critical loads.

The disadvantages of parallel operation are:

1. The generators will be affected by all faults or disturbances on the normal system, thus requiring fast and reliable relaying.

2. The system operators have additional responsibility to synchronize properly the generators and establish desirable reactive and kilowatt loading on all units.

3. Generator stability must be carefully studied and provisions must be made to prevent losing critical loads.

The advantages of separate operation are:

1. Generators are unaffected by disturbances on the normal system.

2. Control of voltage and frequency of the generators is less critical.

3. Generator stability is not affected by faults on the normal system.

The disadvantages of separate operation are:

1. A complicated transfer system is required to transfer critical loads from one source to the other in the event of loss of one of the sources. This could result in critical voltage and torque transients on the generators and motors which might be damaging to them.

2. Control of the normal system voltage and power factor is not possible.

3. A larger number of circuit breakers are required to function during any type of fault to assure continuity of power to essential loads.

Bus Arrangement for Parallel Operation

The 230-kv portion of the station consists of three circuit breakers and two short bus sections in series. A transformer connects to each bus section. The breakers and busses connect in series with a closed 230-kv loop which includes other portions of the plant area. This arrangement is similar to that used in other stations on the loop and no advantage could be found in changing it.

For the 13.8-kv portion of the system, a secondary selective type of system was found to be the most desirable since there are two power transformers and enough load components to divide the total load between the two transformers. This arrangement proved quite satisfactory. The bus tie portion was modified to include a tie-line position which connects to the opposite substation for an alternate source of power when one of the main power transformers is out of service. All loads connected to this portion of the system are interruptible so that no emergency source of power is required for them.

In planning the 4.16-kv portion of the system, a radial type of system was first considered. This would not meet all operating requirements since not more than one primary pump can be lost at a time without jeopardizing plant safety. With a secondary selective type of bus arrangement such as used in the 13.8-kv system, loss of one of the busses would mean losing power to three of the primary pumps. This condition could not be tolerated. A ring-bus type of system was investigated, but this proved to be extremely complicated from the standpoint of relaying and lacked the desired flexibility for arranging loads on the generators or normal system during abnormal operating periods. The reliability of this type of system did not appear to be as good as would be obtained with other arrangements.

A double-bus type of arrangement for critical loads was finally selected. This is the arrangement shown in Fig. 2. All loads requiring emergency power are connected to the double-bus section. Those loads having alternate feeds but still of an emergency nature are connected with single circuit breakers to busses *B* and *C*. Other loads not requiring emergency power, but utilizing 4.16-kv power, are connected to busses *A* and *D*. To maintain continuity of power in the event of a bus fault, an automatic transfer arrangement is provided on each bus in the double-bus section so that operation of the bus differential relay will transfer all critical loads from the faulted bus to the opposite bus. This transfer takes place in about 1 second and is fast enough to prevent loss of head which might cause cavitation of the secondary pumps. With 2-per-cent slip motors, the transient voltage and torque that might be developed because of residual rotor flux will not be serious and, because of the infrequency of bus faults, no appreciable shortening of motor life should result.

In planning how to connect the loads to the various 4.16-kv busses, first consideration was given to having all circuit breakers closed on the operating loads. This plan has the advantage that no transfer scheme is needed to maintain continuity of power following a bus fault. It does have the disadvantages though in that several breakers must open before a bus fault is isolated from the system and reliable backup relaying becomes very difficult to obtain. The reliability of this method of operation seemed less than that of other plans considered so it was rejected.

The operating method finally proposed and accepted is similar to that shown in Fig. 2. With this method of operation, there is only one tie between any two busses. This permits positive or selective tripping for backup protection. The front and back bus sections are connected through the two generator circuit breakers since they are the prime source of power for emergency conditions and the one that should remain intact if all other sources fail. Each transformer supplying power from the normal system connects to only one bus with the bus tie breakers on the opposite busses closed. The arrangement shown in Fig. 2 indicates that breakers *32*, *34*, *39*, *41*, *65*, *67*, *68*, and *70* are closed. A similar arrangement would be to have breakers *34*, *36*, *37*, *39*, *63*, *65*, *70*, and *72* closed. The latter is not preferred since only one source of power would connect directly to each bus supplying building loads that have only siugle breaker feeds.

Fig. 3. A 4.16-kv switchgear

The loads are divided as equally as practicable between the four busses. Only one circuit breaker to each load is closed so that no additional tie points are established between the front and rear busses. The breakers supplying the feeders are interlocked to prevent inadvertent closing of both breakers. With this arrangement of breakers, each bus has two source points. Only two circuit breakers need to open to isolate a bus fault and better selectivity of backup relaying can be obtained than with the other plans considered. With approximately one-fourth of the total load carried on each bus, the least distress to the system will result in the event of a bus failure and a minimum of load must be transferred.

System Components

Transformers

The 230-kv power transformers are oil-insulated with a self-cooled rating of 30,000 kva and a forced-air-cooled rating of 50,000 kva. They have two windings: the primary winding is wye-connected

Fig. 4. A 3,500-kw turbine-generator

for solidly grounded operation and the secondary winding is delta-connected. The primary winding has two $2^1/_2$-per-cent taps above and two $2^1/_2$-per-cent taps below the nominal voltage of 220 kv. The secondary is rated 13.8 kv and has no taps. Nominal impedance is 13.5 per cent.

A zigzag grounding transformer is connected to each 13.8-kv bus. The grounding transformers have an impedance of 1.92 ohms per phase with a 4-ohm resistor connected between neutral and ground.

The 13.8–4.16-kv transformers are oil-insulated and air-cooled with a rating of 5,000 kva. The primary windings are delta-connected and the secondary windings are wye-connected and grounded through the 3-ohm resistor. The 4.16-kv 480-volt transformers are askarel-filled with delta-connected primary windings and wye-connected secondary windings which are solidly grounded.

Circuit Breakers

The 230-kv circuit breakers have an interrupting duty of 5,000 megavolt-amperes (mva) and a continuous current rating of 1,200 amperes. They have an opening time of 3 cycles and a closing time of 20 cycles and are pneumatically operated. The maximum design voltage is 242 kv for use on 230-kv systems with solidly grounded neutrals.

The 13.8-kv circuit breakers are of the metal-clad indoor type, having an interrupting duty of 500 mva. The incoming and bus tie breakers have a continuous current rating of 2,000 amperes and all feeder breakers have a rating of 1,200 amperes. All breakers are solenoid operated with a 125-volt d-c control system. Because of the large induction motor load, it was necessary to specify breakers with a momentary current rating of 60,000 amperes. The normal momentary rating for breakers of this class is 40,000 amperes. The circuit breakers are of the magnetic blast type. The 4.16-kv breakers are the same type as the 13.8-kv breakers but with an interrupting rating of 250 mva and a momentary current rating of 60,000 amperes. Fig. 3 shows a portion of the 4.16-kv switchgear.

Lines and Feeders

The 230-kv lines are of single-circuit construction with flat horizontal configuration and a phase spacing of 27 feet. The conductor is 795,000-circular-mil steel-reinforced aluminum cable, type "drake," with compression-type fittings used for all splices and connections. Two overhead shield wires of 5/8-inch high-strength steel cable and a buried counterpoise of two No. 4 Birmingham Wire Gauge solid galvanized steel conductors provide lightning protection. The counterpoise conductors are buried 18 inches in the ground and emanate radially from each tower to a point 75 feet from the center line on the transmission circuit. All 13.8-kv and 4.16-kv circuits are in underground duct systems in outdoor runs and in expanded metal trays or conduit within the building. Single-conductor shielded cables with butyl ozone-resistant insulation and neoprene jackets are used throughout. All cables are terminated in air with stress-relief cones and have manufacturers' recommended creepage distance over the insulation between the terminal and stress cone.

Generators and Turbines

The generators have a rating of 3,500 kw at 0.7 power factor and a terminal voltage of 4.16 kv. The windings are wye-connected, and each machine is grounded through a 3-ohm resistor. Generator speed is 3,600 rpm. The main and pilot exciters are direct-connected to the generator shaft. The machines are air-cooled and totally enclosed with air-to-water heat exchangers. Class-*A* insulation is used for the armature and class-*B* insulation is used for the fields. The reactive capability of the generators is 80 per cent of rated kilovolt-amperes.

The steam turbines are rated at 3,500 kw with a capability of 125 per-cent of this value, or 4,375 kw with normal steam conditions. The turbines are multistage and condensing with normal steam conditions of 400 pounds per square inch gauge (psig) at 750 degrees Fahrenheit (F) and an exhaust pressure of 2 inches mercury absolute. The turbine and generators are solidly coupled with three bearing-type construction. The steam rate at rated load is about 35,000 pounds per hour. Quick-close flanged-type throttle valves are used for speed control with means for direct mechanical tripping by the overspeed governor and by low oil pressure. Steady-state speed regulation is 4 per cent from no load to full load. The transient loss of speed following sudden application of full load does not exceed 6 per cent and the unit recovers to stable speed within 6 seconds. Fig. 4 shows one of the turbine-generator units.

Boilers

A separate oil-fired steam-generating unit is provided for each turbine-generator set. Each has a maximum rating of 58,000 pounds per hour at a pressure of 415 psig and 750 F maximum. They are designed to operate continuously at a nominal load of 36,000 pounds per hour and be capable of instantly supplying a maximum load of 58,000 pounds per hour with a minimum transient pressure of 360 psig. Electric drives are used on all auxiliaries including feed-water pumps and forced-draft fans. No induced-draft fans are used.

Motors

The raw-water pump motors, rated 1,500 hp, 600 rpm (synchronous), 4.16 kv,

are of vertical squirrel-cage-induction solid-shaft type, and are totally enclosed for outdoor operation with air-to-water heat exchangers. The motors have normal slip and normal torque characteristics. Fig. 5 shows a view of these motors.

The primary pump motors, rated 900 hp, 600 rpm (synchronous), 4.16 kv, are of vertical squirrel-cage-induction solid-shaft and open-dripproof type construction. The rotors have a direct-connected flywheel immediately below the main rotor section to provide a total weight×radius^2 of 77,000 pound-feet^2. Rotor slip at full load is 2 per cent, and the windings are specially designed for across-the-line starting of the high-inertia load.

The secondary pump motors are 10,000-hp, horizontal, wound-rotor induction motors. They operate at 13.8 kv and have a synchronous speed of 600 rpm. Liquid rheostats are provided for secondary control of the motors for starting the high-inertia-connected load and for speed control to regulate the flow of reactor cooling water. A circuit breaker is provided to short-circuit the liquid rheostat during periods of full-speed operation. The motors are totally enclosed with air-to-water heat exchangers. The complete secondary pump drive unit consists of the motor, speed increaser unit, a forged flywheel, and pump. The total inertia of the system referred to the pump shaft is 156,000 pound-feet^2. Fig. 6 shows these units with the primary pump motor in the background.

All other motors over 200 hp are squirrel-cage induction motors operating at 4.16 kv and with normal torque characteristics. They are of the open-dripproof type construction. All motors of 200 hp and smaller operate at 480 volts. All the afore-mentioned motors employ full voltage starting.

A-C Network Analyzer Study

To determine the limits of stability and the necessary relaying required for protection of the system, extensive studies were performed of the entire electric system on an a-c network analyzer. These studies showed that with a 3-phase fault anywhere on the system the generators would remain stable for a fault time of 20 cycles. If the fault persisted beyond this time, instability might occur. A detailed check of the BPA system was not made for stability, but indications were that no problem would exist within the 20-cycle period for faults near the plant being studied. The first parts of the system to become unstable were

Fig. 5. Raw water pump motors

the large induction motors. This was due to low voltage caused by the high accelerating current of all motors following removal of the fault and which resulted in insufficient torque to permit the motors to accelerate to normal speed.

Because of the predominant induction motor load on this system, a complete equivalent circuit representation of the large induction motors was made on the board.[3] Values of R_s/S were estimated for each step, and these values were inserted in the board so as to give a true representation of the action of the motor during faults when slip would be changing. This was also important in studying the acceleration of the motors following clearing of the fault because of the high-inertia load and low-voltage conditions that would exist while the motors were attempting to accelerate. The conventional method of representing induction motors by a fixed impedance or by a fixed kilowatt load would have given completely erroneous results in this case.

Grounding

The various system voltages are grounded in accordance with the latest AIEE recommendations.[4] This was done to prevent transient overvoltages which could damage or shorten the life of cables, motors, and other equipment. It also provides positive ground relaying and immediate isolation of faulted circuits.

The primary windings of 230-kv transformers are wye-connected and solidly grounded. This conforms to the method of grounding used by BPA on their 230-kv system.

The 13.8-kv system is grounded by means of zigzag grounding transformers with 4-ohm neutral resistors. Each transformer has a grounding transformer and resistor which give a maximum grounding fault current per transformer of 2,000 amperes. This method of grounding the 13.8-kv system was used in other locations of the plant and appeared to be satisfactory for use in this new area.

There was no previous experience at Hanford on 4.16-kv systems, so various methods of grounding were studied. Because of the generators connected to this system, solid grounding of the generator neutrals could not be used since the generator current for ground faults would exceed the maximum 3-phase fault current. Reactance grounding was not considered satisfactory since transient overvoltages could still exist under certain conditions. Low-value reactance grounding would satisfactorily limit ground current to 3-phase values but the high-fault currents would result in severe burning of core iron in the equipment.

Resistance grounding appeared to offer the best solution for this system with its many possible methods of operation. Normally, two generators and two transformers will be paralleled on a group of busses; however, there may be instances when any single transformer or generator may be isolated from the system and supply considerable load. It was necessary that each source of power be capable of providing ground current and that no special switching be required when a generator or transformer is added or removed from the system. Employing a resistor in the neutral of each transformer and generator was explored. It was necessary that adequate ground current be available for satisfactory relay operation when only one source of power was operating. This condition was satisfied by the use of a resistor giving a ground current of 800 amperes, or approximately full load current for each transformer or generator. With four sources in parallel, this would

Fig. 6. Secondary pumps and drives; primary pumps are in background

result in a total ground current of 3,200 amperes. This is just over 10 per cent of total 3-phase fault current and is well within the 5- to 20-per-cent range recommended by the AIEE. These values are adequate to give satisfactory relaying and still keep burning at the fault to a minimum.

In the final design, individual 800-ampere resistors were permanently connected in the neutral of each generator and 4.16-kv transformer winding. Satisfactory grounding is provided with any conceivable switching arrangement when not more than four sources of power are in parallel. Since five or more sources of power would exceed the interrupting rating of the 4.16-kv switchgear, an operating limit of four sources has been established.

Relaying

Lines and Feeders

The 230-kv lines are protected by power directional comparison relays with carrier control. Impedance-type phase relays with three zones of protection are used for backup protection against phase faults, and product-type ground relays are used for backup protection of ground faults. The ground relays are polarized by residual voltage obtained from line potential devices. A backup time-delay tripping relay is provided to trip the 230-kv tie circuit breaker in the event a line circuit breaker fails to open following operation of the line protective relays.

The 13.8-kv tie line feeder is protected by pilot-wire relays with backup protection provided by phase and ground overcurrent relays with inverse-time characteristics. The feeder lines on the 4.16-kv system have phase and ground overcurrent relays with inverse-time characteristics and instantaneous overcurrent elements for phase-to-phase short-circuit protection.

Transformers

All power transformers having a primary voltage of 230 kv or 13.8 kv are protected by differential relays having harmonic restraint. The 4.16-kv 480-volt transformers for station service power are also differentially protected. All other transformers have time overcurrent relays with inverse-time characteristics for phase and ground fault protection and instantaneous elements for phase short-circuit protection.

The inverse-time relays are set to give backup protection to the transformer secondary relays and the instantaneous elements are set to protect for faults into about 80 per cent of the transformer impedance. The 13.8-kv zigzag grounding transformers are outside the transformer differentially protected grounding zone and so require separate relaying. Inverse-time overcurrent relays are provided in each phase lead for protection against internal transformer faults. The current transformers for these relays are delta-connected, which permits high sensitivity for internal transformer faults resulting in a flow of positive- and negative-sequence current and makes the relay insensitive to a flow of zero-sequence current which would result during an external ground fault.

Busses

All 230-kv, 13.8-kv, and 4.16-kv busses are protected by high-speed current-differential relays of the high-impedance voltage-actuated type. Backup protection for the 13.8-kv busses as well as for the 13.8-kv feeders is provided by inverse-time overcurrent relays on incoming 13.8-kv circuit breakers *1* and *2*. Backup relaying for phase faults on the 4.16-kv busses is provided by inverse-time overcurrent relays on the primary of the 13.8–4.16-kv transformers supplying the busses. Special backup protection for ground faults on these busses is covered in a later section. One exception to the normal method of bus protection was made with the 13.8-kv tie section which includes breakers *8*, *9*, and *10*. Rather than provide a separate set of bus differential relays for this section it was included in the protective zone for bus *1*. Since these busses are not included in the emergency system, no hazard to the plant would result if both bus sections were lost following a relay operation. The additional cost of providing extra relays and current transformers therefore did not seem justified. Inverse-time overcurrent relays are also provided on each bus tie position for backup protection when the tie circuit breakers are closed.

Generators

Current differential protection is provided on each generator for primary relaying. For backup protection for external overloads, inverse-time overcurrent relays with voltage restraint are provided. Residual current relays are included in both the generator breaker current transformers and the generator neutral. Co-ordination of these relays is covered in a later section, with relays to detect loss of field excitation included.

Motors

The 10,000-hp wound-rotor induction motors are protected by current differential relays, primary current balance relays, thermal relays for overload protection, instantaneous overcurrent relays for short-circuit protection, and inverse-time overcurrent relays for ground fault protection.

The 1,500-hp raw water pump motors have current balance relays, thermal relays for overcurrent protection, instantaneous overcurrent relays for short-circuit protection, and inverse-time overcurrent relays for ground fault protection. All other 4.16-kv motors have thermal overcurrent, short-circuit, and ground fault protection. All 480-volt

motors have conventional-type magnetic starters with thermal overloads and fuses or breakers for short-circuit protection. Some of the larger size 480-volt motors use air circuit breakers with suitable series-type overload devices.

Special Backup Relays

The first line protection for motors, feeders, transformers, busses, and other devices presented no particular problem since standard methods of protection could be readily applied. The backup system, however, proved to be quite complicated, particularly on the 4.16-kv system with its six busses. Since the system is grounded and all outgoing circuits are of single-conductor shielded cable, the predominant type of faults expected are the single-line-to-ground type. For this reason, special emphasis was placed on backup relaying for ground faults.

Backup relaying is required to clear a fault in case a circuit breaker or first line relay fails to function properly. If this occurs, the next place to isolate or clear the fault with the least distress to the system would be the bus supplying the fault and its source breakers. The first approach to provide backup protection was through the application of time overcurrent relays on the source breakers for each bus. It soon became evident that proper co-ordination could not be obtained without introducing extremely long time delays in certain of the relays. Since this portion of the system supplies the critical shutdown and emergency loads, the fastest protection obtainable is desired to assure safety to the plant.

Table I. Fault Currents

	Fault on Bus	Current in Breaker, Amperes						
		32	34	39	41	65	67/68	70
Case I	*B*	*800	*2,400	800	800	1,600	1,600	1,600
	E	800	800	800	*1,600	*1,600	*1,600	1,600
	C	800	800	*2,400	*800	1,600	1,600	1,600
	F	800	800	800	800	1,600	*1,600	*1,600
		34	36/37	39	63	65	70	72
Case II	*B*	*1,600	*1,600	1,600	800	800	800	800
	E	1,600	1,600	1,600	*800	*2,400	800	800
	C	1,600	*1,600	*1,600	800	800	800	800
	F	1,600	1,600	1,600	800	800	*800	*800

* Values are for circuit breakers directly supplying the faulted bus.

Table I shows fault currents that would flow through various circuit breakers for a fault on each bus or a feeder connected to that bus. In case I of this table it is assumed that tie breakers *67* and *68* are closed between busses *E* and *F*, transformer *2* is supplying bus *B*, generator *1* is supplying busses *B* and *E*, transformer *3* is supplying bus *C*, and generator *3*, is supplying busses *C* and *F*. The fault current values assume a current of 800 amperes from each source, which will be approximately true, if zero fault resistance is assumed. The values marked by an asterisk are for the circuit breakers directly supplying the faulted bus and the ones that should open first to isolate the fault without causing separation of the rest of the system. For a fault on bus *B*, it is noted that breakers *32* and *34* should open ahead of other source breakers. Breaker *34* has a current of 2,400 amperes so no problem should exist in setting its relays. Breaker *32*, however, carries only an 800-ampere fault current and must trip while other breakers in this system with equal or higher current must not trip.

Look now at the condition for a fault on bus *E*. Breakers *65*, *67*, and *68* must trip first with each breaker carrying a fault current of 1,600 amperes. These breakers and also breaker *32* carry the

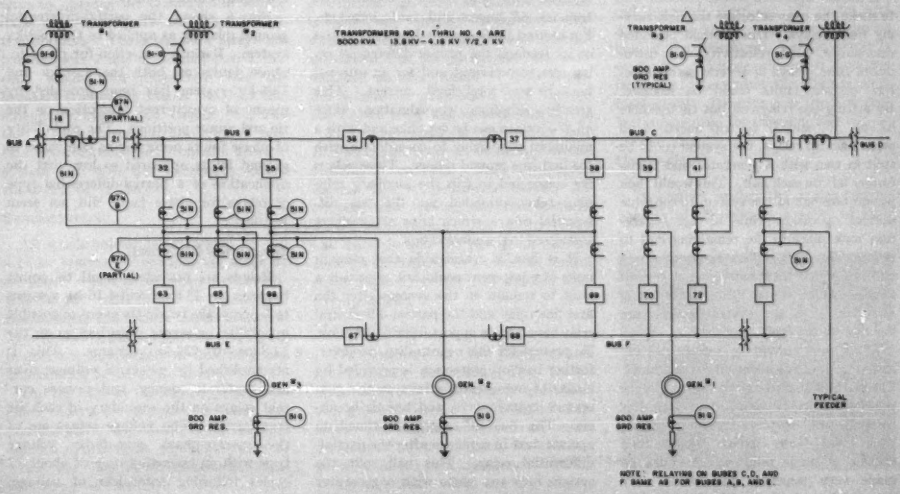


Fig. 7. Residual relaying, 4.16-kv system

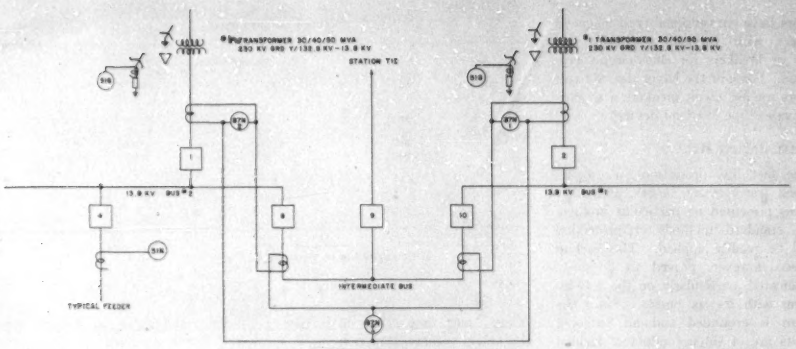


Fig. 8. Residual relaying, 13.8-kv system

same current as for a fault on bus *B*, but now the opposite sequence of tripping is desired. The use of directional relays would not solve the problem since direction of current flow for some breakers is the same for a fault on any bus. Breaker *70*, for example, carries a 1,600-ampere fault current for a fault on bus *F* and should trip, but it also carries the same fault current in the same direction for a fault on bus *B* and should not trip. Case II shows how the currents would divide with tie circuit breakers *36* and *37* closed and circuit breakers *67* and *68* open.

A second approach to the problem was to make the relay selective through varying time delays. This method provided reasonably good selectivity but introduced time delays of several seconds before certain faults could be isolated. By setting the relays on bus tie breakers *67*, *68*, *36*, and *37* to trip ahead of all other backup relays, the system could be split in two with a generator and transformer left on each half. This would then permit one half of the system to continue normal operation while further separation took place in the remaining half to isolate the fault. This method would still result in a total fault time of several seconds which would cause considerable disturbance to the system and severe burning at the fault.

A method providing faster clearing and more positive selectivity was desired. The method finally devised was to totalize the fault current supplied to each bus through all source circuit breakers. This total fault current would then operate a single relay which could be made very sensitive and would only have to be selective with the first line relay on the feeder breakers. This relay has been called a partial-differential type of relay. A schematic of this relay system is shown in Fig. 7. On bus *B*, for example, the current in breakers *21*, *32*, *34*, *35*, and *36* is totalized and connected to a residual relay called an *87N*. Referring again to Table I, the total current seen by this relay would be the vector sum of the currents in breakers *32* and *34* or a total of 3,200 amperes. For a fault on bus *E*, however, breakers *32* and *34* each carry 800 amperes, but the direction of current in breaker *34* is now reversed so that the total relay current is zero. This same result is found to be true for all busses and fault locations. For ground faults external to a given bus or its feeders, the partial-differential relay sees zero current and for an internal fault it sees total fault current. This greatly simplifies co-ordination since all *87N* relays can be set alike and with a minimum time delay to co-ordinate with the first line ground relays. These relays are connected to trip the auxiliary tripping relay provided for the bus differential relays, which trips all breakers connected to the affected bus.

It is hardly conceivable that enough units of equipment could fail to permit a fault to remain on the system after the first line relay and the partial-differential relay have had an opportunity to function. To protect for this eventuality, however, further backup protection is provided by means of overcurrent relays on the generators, transformers, and bus tie breakers. The bus tie relays are timed to operate next in sequence after the partial-differential relay. This will split the system into two parts with a generator and transformer supplying each. The next step is to trip the generator breaker nearest to the faulted section. This then separates the two busses in the faulted half of the system. If the fault is on the bus section supplied by a transformer, then the transformer circuit breaker will operate after the generator breaker to clear finally the fault. The transformer circuit breaker is given the longest time delay so that it will not trip ahead of the generator breaker if the fault is in the opposite bus section. Should all these devices fail, a last backup is provided in the neutrals of the generators and transformers to trip their respective differential auxiliary relays.

Fig. 8 shows the partial-differential ground relaying as applied to the 13.8-kv system. Backup protection for phase-to-phase faults on both the 4.16-kv and 13.8-kv system has been provided by means of overcurrent protection on the tie and source positions. The probability of phase faults occurring as compared to ground faults appeared so low that the application of a partial-differential type of relays for phase faults did not seem justified.

Separation Relaying

Relays are provided on all tie points between the 13.8-kv and 4.16-kv systems to separate the two in the event of possible instability or severe disturbances on the 13.8-kv or 230-kv systems. This is accomplished by means of voltage, over and under frequency, and reverse current relays on the secondary of each tie transformer. The voltage relays are of the reverse-phase over-under voltage type with an operating time of about 12 cycles following total loss of voltage. The frequency relays have settings that

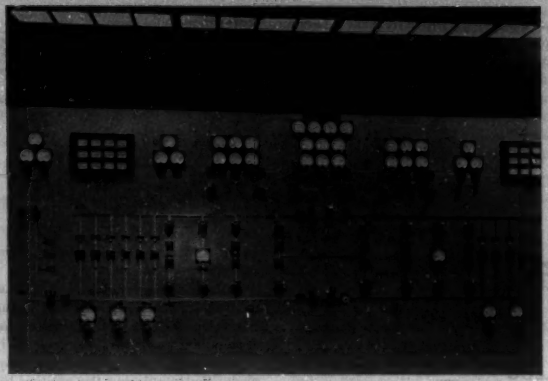


Fig. 9. A 4.16-kv control board

will permit the normal fluctuation of normal system frequency to occur without separation of the two systems. These relays have inverse-time characteristics and will permit system operation between about 57 and 62 cycles. The reverse current relays permit a nominal power flow from the 4.16-kv system to the 13.8-kv system but will operate during instability or abnormal conditions. One set of relays is provided on each transformer secondary. Each set trips a common hand reset auxiliary relay to cause separation of the normal and emergency systems by opening the primary and secondary circuit breakers on the affected transformers. The separation relays on transformers *1* and *4* are inoperative except when their respective bus tie breakers connecting bus *A* or *D* to the emergency system are closed. Separation of the two systems also trips off certain nonessential loads from the 4.16-kv system but does not necessarily cause shutdown of production.

Synchronizing

All synchronizing is performed on the 4.16-kv system. Provisions are made at all points where an asynchronous condition might exist such as generator, transformer, and bus tie circuit breakers. The double-breaker positions supplying motors are interlocked so that only one breaker can be closed at a time.

On the 230-kv system, synchronism check relays are provided on each line position permitting closing of the breakers when a synchronous condition exists or when the line is de-energized. Synchronizing of separate sources at this point is not considered necessary.

Control

Location and Arrangement of Controls

All circuit-breaker and motor controls are located in a central control room with the exception of the 230-kv controls and the incoming and tie breaker positions on the 13.8-kv system. In the central control room there are two main control boards. One is a graphic panel type of control with all pump motor controls and the necessary instrumentation required for the water plant. The second control board includes all 4.16-kv breaker controls not directly associated with water pumping. Control and instrumentation for the generators are also included on this board. Fig. 9 shows the 4.16-kv control board and Fig. 10 the 230-kv relay and control board.

Protective relays for the 4.16-kv and 13.8-kv circuit breakers are located on the switchgear cubicles. Each switchgear cubicle has a permissive switch for tripping and locking out the breakers during maintenance. The 230-kv relays and controls and the remaining 13.8-kv controls are located in a separate control room adjacent to the two outdoor 230-kv substations.

Motor Control

The majority of the pump motors are manually controlled. The 10,000-hp wound-rotor motors have individual and group control on their liquid rheostats. The primary motor circuit breakers cannot be closed until the rheostat is in a 37-per-cent resistance position which will limit initial starting current to 125 per cent of full load current. The rheostats return to this position automatically after the motor is shut down. After starting the motor, the rheostat is adjusted manually to the desired running speed. If several motors are to be started, each motor is started individually; then the group can be accelerated together by means of a master control on the liquid rheostats. Individual controls are then used for balancing flow between the pumps. If the motors are to operate at maximum speed, the liquid rheostats are adjusted for minimum resistance; then a short-circuiting breaker is closed which short-circuits the residual resistance.

Annunciators and Alarms

Visual and audible annunciation is provided on all circuit breakers and important relays to notify the operator of system disturbances. In general, the amount of information given to the operator is determined by the necessary action required of him. If the trouble is of such a nature that the operator must take immediate action to shut down a piece of equipment or operate a circuit breaker, special information as to the type of trouble is made known to the operator. If it is the type of warning that calls for a floor man to check on such things, as water flow or temperature of a certain machine, then the operator is only cautioned that a machine needs attention.

Station Batteries

There are two sets of station-type glass cell batteries rated at 120 volts and 320 ampere-hours. Each set supplies half of the switchgear, but provision is made for supplying all circuit breakers from one battery when the other is being maintained. Diverter-pole type of generators are used for charging them.

Conclusion

To confirm further the design and reliability of the system, a series of fault

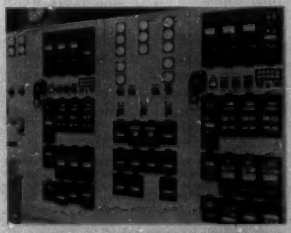


Fig. 10. A 230-kv control board

tests were performed. These tests consisted of placing 3-phase and single-phase-to-ground faults at critical points on the system. No evidence of instability was noted during any of the tests and, where separation of the normal and emergency systems occurred, the generators picked up full load with less than a 6-per-cent dip in frequency. Some problems in relaying were encountered but these have been corrected. The relatively short period of time that this plant has been in operation indicates that the basic concept of design is sound and that reliability of power to the emergency loads is of a high order.

References

1. MAKING ATOMIC PILES BEHAVE, Ivan M. A. Garcia. *General Electric Review*, Schenectady, N. Y., Mar. 1955, pp. 45-49.
2. POWER SYSTEM STABILITY (book), S. B. Crary. Vols. I, II, John Wiley & Sons, Inc., New York, N. Y.
3. GROUNDING OF INDUSTRIAL POWER SYSTEMS, *Publication 953*, AIEE Standards series, 1956.

A Time-Dependent Nonlinear Compensating Network

JOHN C. CLEGG
ASSOCIATE MEMBER AIEE

Synopsis: A network for improving the transient response of a-c or d-c servomechanisms is described. Variations of the carrier frequency in a-c systems have no effect on the operation, and the magnitude of the step function input may be varied over wide limits without need for adjustment to obtain optimum response. The components used are reliable, few in number, and readily obtainable.

THE NETWORK under consideration is designed for placement in the error path in the forward loop of the servomechanism to be compensated as shown in Fig. 1. As can be seen in Fig. 2, the basic compensating network consists of two diodes, two capacitors, and three resistors. Except for diode polarities, the two branches of the network are identical; that is $C_A = C_B$ and $R_A = R_B$. It is assumed that the input resistance to the following stage is combined with R_0.

A complete mathematical analysis of the effect of this network in a closed-loop system is difficult to make because the characteristics vary with both time and amplitude. Fortunately there are only two important design parameters (the time constants $C_A R_A$ and $C_A R_0$) which must be adjusted, and so a trial-and-error synthesis is not unreasonable.

Qualitatively, the operation of the network can be explained by the following reasoning. Assume both capacitors to be initially discharged, and assume a positive step function voltage to be applied at R (Fig. 1), the reference input. The output quantity C cannot respond instantly, so an error voltage E is instantaneously applied to the compensating network. Since E is positive, the upper diode of Fig. 2 conducts, and because capacitor C_A cannot acquire a charge instantaneously, the entire error voltage initially appears across R_0 at the network output. This compensating network output voltage F causes the remainder of the servo system to make a correction in the controlled quantity C such that the error begins to diminish.

As time progresses, C_A charges with a time constant of approximately $C_A R_0$ (assuming $R_A >> R_0$) which is made somewhat shorter than the natural period of oscillation of the servomechanism. As soon as the rising voltage of C_A equals the falling error voltage, the diode stops conducting, thus opening the signal path and allowing the controlled quantity C to coast toward the zero error point with no further driving force. Ideally, inherent damping, which is the only force active during the coasting period, would cause the controlled quantity to stop at or near the zero error point.

If the controlled quantity should stop before reaching the zero error position, it would not remain stopped long, for when C_A discharges sufficiently through R_A, the upper diode can again conduct, reclosing the signal circuit and causing a further reduction of error. This process can continue indefinitely if necessary.

For negative step function inputs, C_B, R_B, and the lower diode are active instead of the upper branch of the network. If there is an overshoot in the output such that the error is first positive and then negative, first one branch of the network and then the other is active.

In many servomechanisms, the error appears as a modulated alternating voltage. The behavior of the network toward such a voltage is essentially the same as has been described except that the upper diode conducts during the positive half-cycle of error voltage and the lower diode conducts during the negative half-cycle. The network as before has an initial period of conduction followed by a dead period during which the controlled quantity may coast toward the zero error position. Variations in carrier frequency do not affect the response.

The switching point where one of the diodes stops conducting is determined by the voltage on its associated capacitor, and this voltage depends on the magnitude of the step function input. The network consequently, when adjusted to optimum response for one magnitude of input, will have very nearly optimum response for all magnitudes if the remainder of the servomechanism is reasonably linear.

A simple method of providing increased damping is shown in Fig. 3 where a parallel path through a sign-changing amplifier or transformer is added around the original network. (This sign-changing amplifier or transformer serves mainly to reverse the polarity of the error voltage, but it may have in addition a voltage gain or loss.) As long as either of the diodes is conducting, the sign changer is relatively ineffective because the impedance of the diodes and capacitors is made much lower than the resistance of R. However, when the diodes are not conducting, the sign changer applies to the output terminal of the network a voltage which is opposite in polarity to the error voltage, thus causing the output torque to reverse, decelerating the controlled quantity more rapidly than could be done by inherent damping alone. The system would now be unstable (since the negative error signal tends to drive the controlled quantity away from the zero error position) if it were not for the fact that the capacitors discharge at about the same

Paper 56-778, recommended by the AIEE Feedback Control Systems Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Summer and Pacific General Meeting, San Francisco, Calif., June 25-29, 1956. Manuscript submitted July 18, 1955; made available for printing April 26, 1956.

JOHN C. CLEGG is with the University of Utah, Salt Lake City, Utah.

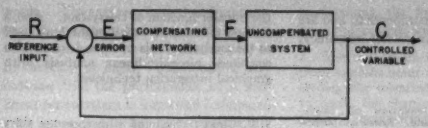


Fig. 1. Relation of compensating network to other system components

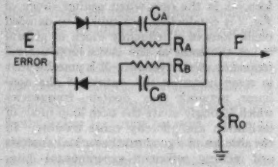


Fig. 2. Basic circuit for compensating network

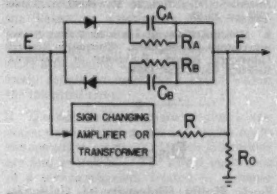


Fig. 3. Compensating network with added damping

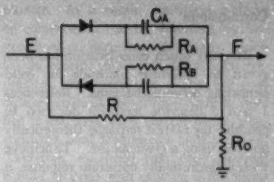


Fig. 4. Compensating network having reduced sensitivity to noise

rate that the error is reduced. Thus, by the time the controlled quantity has reached the null point, the capacitors are essentially discharged, allowing the diodes to conduct if necessary to override any negative error signal from the sign changer.

Noise superimposed on the error voltage will be rectified and stored on the capacitors causing the networks of Figs. 2 and 3 to be insensitive to error voltages smaller than the noise. This may or may not be a serious limitation depending on the accuracy requirements, the noise magnitude and wave shape, the feasibility of shielding and filtering, etc.

The basic network can be made less sensitive to noise by adding a parallel resistor R as shown in Fig. 4. Following a step function input, this network will initially have a gain of unity, causing a relatively large correcting force; but before the error reaches zero, the diodes stop conducting, and the gain is reduced to $R_0/(R+R_0)$. Most servomechanisms tend to be more stable with reduced gain; therefore, during the period of reduced gain, the transient terms are allowed to die out quickly. As the capacitors discharge, full gain and steady-state stiffness are restored to the system. The only effect of noise on this network is to reduce the gain at low signal amplitudes.

Experimental Results

Analogue computer studies have been made which demonstrate the usefulness of the proposed basic compensating network and its variations. The oscillograms shown in Figs. 5 and 6 were obtained from a simulated third-order servomechanism having the transfer function

$$KG(s)=\frac{5}{s(0.6s+1)(0.1s+1)} \qquad (1)$$

Curve A of Fig. 5 shows how the controlled variable C responds to a step function input for the uncompensated system. Curve B shows the response of the same system after adding the compensating network of Fig. 3* and increasing the gain. Notice that there is no overshoot and that the rise time is shorter. For comparison, curve C shows the response of the uncompensated system after the gain has been reduced to eliminate overshoot.

*The network components used in obtaining these curves were as follows: $R_A=R_B=5$ megohms, $R_0=0.25$ megohm, $R=13$ megohms, $C_A=C_B=0.05$ microfarad. A 6H6 vacuum duodiode was used, and the sign-changing amplifier had a voltage gain of 2.

The oscillograms of Fig. 6 show the error following a ramp input for the same two systems as curves A and B in Fig. 5. Again curve A represents the uncompensated system and curve B the compensated system with increased gain. It can be seen that the transient terms in the error of the compensated system are of smaller magnitude and shorter duration and that the steady-state velocity error is smaller for the compensated system. The network of Fig. 3 has also been used successfully in the a-c error path of a practical fourth-order magnetic clutch servomechanism.

Nonlinear Systems

A large amount of saturation was introduced into the compensated system, and it was found that the only noticeable effect on the step function response was an increase in response time, there still being no overshoot. The response time of the uncompensated system with the same amount of saturation was found to be even longer, which demonstrates that the network is not restricted to use with linear systems. In fact, the network of Fig. 2 is effective in suppressing small amplitude oscillations or jitter caused by backlash. Any tendency to oscillate causes the network to introduce a temporary dead zone for small amplitudes into the forward loop. Backlash oscillations cannot be sustained in a loop containing such a dead zone; therefore oscillations cease.

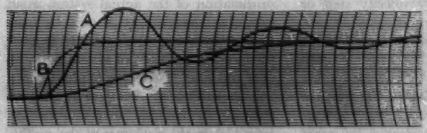


Fig. 5. Response C(t) to step function input

A—Uncompensated system with normal damping
B—Compensated system with increased gain
C—Uncompensated system critically damped
Time scale: 5 lines=1 second

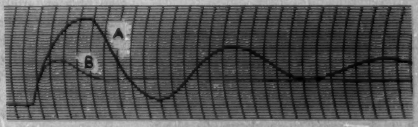


Fig. 6. Error E(t) for ramp input

A—Uncompensated system with normal damping
B—Compensated system with increased gain
Time scale: 5 lines=1 second

Conclusions

Recently much consideration has been given to nonlinear methods for compensating servomechanisms. References 1 and 2 describe methods designed to produce the fastest response theoretically possible using a given motor. To achieve such a theoretically optimum response, a rather complex computer is required to determine the correct time for reversing the motor torque. The time-dependent nonlinear network presented in this paper, while not capable of producing the theoretically fastest response, can cause a significant improvement in the performance of many servomechanisms without particularly complicating the circuitry.

One of the important features of the proposed network is the fact that it can be used with modulated a-c error voltages. This fact may often make possible the elimination of tachometers or demodulators with their expense, time delay, and filtering problems, or make possible the elimination of a-c lead networks such as the bridged T or parallel T networks which require nonstandard components and are sensitive to changes in carrier frequency.

References

1. Multiple Mode Operation of Servomechanisms, D. C. McDonald. *Review of Scientific Instruments*, New York, N. Y., vol. 23, no. 1, Jan. 1952, pp. 22–30.

2. Nonlinear Optimization of Relay Servomechanisms, L. M. Silva. Electronics Research Laboratory Report, University of California, Berkeley, Calif., Apr. 1954.

Discussion

F. J. Mullin (University of California, Berkeley, Calif.): This paper presents a straightforward method for compensating a-c or d-c servomechanisms. However, in his paper the author uses a transfer function which has its poles restricted to the negative real axis, and as a result of this assumption the question arises as to whether this method of compensation is equally applicable to systems which have complex poles. This point was not discussed in the paper and I would like to know if this method of compensation tacitly places some restrictions on the transfer function of the system. Also, a method of solution which might lend itself to the analysis of this problem is one which combines piecewise-linear analysis with graphical integration techniques.

J. C. Clegg: Systems with complex poles in the left half plane when operating open loop, as is the case when neither diode of Fig. 2 is conducting, are subject to damped oscillations of a frequency which in general is not the same as the closed-loop natural frequency of the system. It is possible that in some instances the output of the nonlinear network may contain frequencies which strongly excite the open loop mode of oscillation and thereby cause trouble. In the absence of a good mathematical analysis and lacking sufficient experimental data, I am reluctant to make a general statement; but it is my belief that this network may be used to improve the stability of a great many systems which would be benefited by a phase lead regardless of whether the poles are real or complex. It is true that a piecewise-linear analysis can be made of this network, but the procedure is very cumbersome. The time (or voltage) at which the diodes switch is not known beforehand and boundaries of the two linear regions must be determined by a series of successive approximations.

An Experimental Treatment of Nonlinear Servomechanisms

R. S. NEISWANDER
ASSOCIATE MEMBER AIEE

AN EXPERIMENTAL method of achieving appropriate first switching points for certain nonlinear single-loop servomechanisms is described. The servomechanism is assumed to be a positioning device. For control intelligence, nonlinear corrective functions of output velocity, static error, or error velocity are assumed continuously available. Since the method is experimental, either the actual device or a reasonably accurate analogue model can be used, and thus is not subjected to the same limitations as those imposed by simplified mathematical models. Here the basic device may be of high order and include many unintentional nonlinearities. One principal limitation accepted by the method is that the system input can be described by a limited Taylor series; i.e., at time $t = 0$, the quiescent system is confronted with an input consisting of an arbitrary position step plus an arbitrary velocity step, or ramp, plus an arbitrary acceleration step, etc. Switching points subsequent to the first are assumed to be compacted into the terminal region of the response, either to be absorbed by allowing the terminal region to be unsaturated, and perhaps linear, or to be determined by other methods. This assumption might be interpreted in practice as the requirement that the system components such as amplifiers, actuators, and sensing elements be in themselves relatively well damped. A highly underdamped element that tends to wiggle or ring during the positioning operation might well require a more elaborate experimental treatment.

Paper 56-777, recommended by the AIEE Feedback Control Systems Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Summer and Pacific General Meeting, San Francisco, Calif., June 25–29, 1956. Manuscript submitted August 8, 1955; made available for printing June 15, 1956.

R. S. Neiswander is with the Hughes Aircraft Company, Culver City, Calif.

This paper is a summary of portions of a thesis investigation conducted during 1953 and early 1954 at the California Institute of Technology under the auspices of the Howard Hughes Fellowship. Dr. R. H. MacNeal was responsible for guidance and stimulation of the work, which was an outgrowth of investigations previously reported.

The Experimental Approach

Real-life servomechanisms are intrinsically complicated. Amplifiers saturate and electric and hydraulic actuators are not linear, and the neat dispatching of an instability by a clever network simply creates several new problems when the new stability limit is pushed. Graphical or analytic methods become tedious and eventually experimental methods have to be used. This transition to an experimental technique has one fundamentally important consequence: the system itself need no longer be approximate. An approximation of the actual system input somewhat resembling that used by more elementary treatments is retained to allow the formation of rules of system performance and rules of effecting improvements. The experimental technique cannot be couched in rigorous mathematics for, indeed, it is the weakness of the mathematics that has demanded the experimental tests.

The beautiful consistency of analytic techniques is not evident in nonlinear systems. Part of the difficulty is with

words. The engineer finds that frequency analysis and step-function analysis of a control system give results that appear to be directly at odds. One method may indicate that the performance of a nonlinear servosystem is improved (compared with linear standards) by decreasing rate feedback and another method may indicate that the rate feedback should be increased to obtain good performance. No methods agree, but all are correct.

The first difficulty is a semantic one. The class "nonlinear servomechanisms" is nebulous—it encompasses all except perfectly linear servomechanisms. It is an example of dichotomy pleasing to the mathematician and frustrating to the engineer. A servomechanism is, by agreement, a device which exerts a fairly strong influence on an environment, or load, controlled by a rather weak request or input. Furthermore, the servomechanism takes advantage of a comparison between its output and its input to effect good control. Best control (synonymously, "optimum performance" or "perfect responses") is thus desirable; here agreement ceases and specific definitions must be admitted. First, this requires a definition of the input to the system of interest; a servomechanism designed to work well from sampled input information would probably work very poorly if a new input were introduced consisting of a continuous signal accompanied by considerable noise. Second, care must be exercised in the use of the words "best" or "optimum" or "perfect." The output of a real system will always deviate from its requested value, and the manner in which this deviation or error is assessed establishes the index of the system's performance. Furthermore, in the design of a servomechanism, the engineer should have some "best" or "ideal" system in mind as a criterion. It is interesting to note that the customary linear servo theory implies an unusually frustrating ideal, a system with infinitely rapid responses. The term "linear" is used here in the engineering sense, referring to systems which can be described by ordinary, constant-coefficient linear-differential equations.

Operating Conditions

A common characteristic of a nonlinear system is the variation of its responses with different operating conditions. A system may be stable for small excitations and unstable for large, or vice versa. Therefore, unlike linear treatments, nonlinear analyses require that the operating conditions be defined. For some of the commonly encountered servomechanisms, the following allocation of operating regions suffices:

1. The null region is the region in which small motions occur without excursions exceeding the tolerated error of the system. In itself, the character of the motions is generally unimportant as long as the tolerance limits are not exceeded. Stability in the linear or asymptotic linear sense, Liapounoff stability,[2] is unimportant; a system might well be unstable about its exact null point if a stable limit cycle (perfectly repetitive oscillation) exists within the tolerated error.

2. The optional unsaturated region is the region extending from the null region up to the operating boundary at which saturations are first evidenced. The useful operations of a servomechanism treated as a linear system are entirely restricted to this region. Contractor (relay) control systems eliminate the region. Within the region, the system may be nearly linear or may have intentional and unintentional nonlinearities.

3. The saturated region is the zone extending as far as desired beyond the saturation boundary.

Performance Description

The description of an input to the servomechanism as a general function $x_i(t)$ is not a useful form in either linear or nonlinear analysis. Usually the description is reworded as a series; e.g.

1. Fourier series (or in the limit, Fourier integral)

$$x_i(t) = a_0 + b_1 \cos \omega t + c_1 \sin \omega t + b_2 \cos 2\omega t + c_2 \sin 2\omega t \ldots$$

2. Taylor series

$$x_i(t) = x_{i_0} + \dot{x}_{i_0} t + 1/2 \ddot{x}_{i_0} t^2 \ldots$$

3. Power series

$$x_i(t) = a_0 + a_1 t + a_2 t^2 \ldots$$

Various frequency methods are based upon a Fourier representation of the input. If the system is linear, superposition is always permitted; responses to a single-frequency input for all practicable values of this frequency describe the system. Now some of the characteristics of a nonlinear system will be considered: 1. superposition is not usually acceptable; 2. a single-frequency input f produces a distorted output having frequencies f, $2f$, $3f$ and, in some instances, subharmonics $f/2$, and $f/3$; and 3. a multiple-frequency input such as f_1 and f_2 produces modulation or beat frequencies such as f_1+f_2 and f_1-f_2. Extended linear-frequency analysis, proposed by Kryloff and Bogoliuboff,[3] assumes a single-frequency input and ignores all but the fundamental frequency of the distorted output. As pointed out by Kochenburger,[4] this can in part be justified in servomechanisms which, being most responsive to low frequencies, tend to suppress the higher harmonics. As a result, nonlinear systems or components can be described by impedances or describing functions which are in general dependent upon both frequency and amplitude. The low-pass characteristic of the servomechanism does not reject, and, as a matter of fact, encourages, beat frequencies such as f_1-f_2. Therefore, extended frequency analysis should be used with caution if the input cannot be well described by a single frequency, and system nonlinearities are sufficiently prominent to cause beat frequencies or large harmonics. It is also noted that extended linear-frequency analysis supports the same "ideal" systems as linear analysis, i.e., those having infinitely fast responses.

Description of the input by a Taylor series is closely related to transient analysis. At time $t=0$, the system is subjected to a displacement step function x_{i_0}, plus a velocity step (ramp $\dot{x}_{i_0}t$), plus an acceleration step $1/2\ \ddot{x}_{i_0}t^2$, etc. After the output has settled, a new transient may be initiated. In this fashion an input function can be approximated by a sequence of transient forcing functions. Since the system is nonlinear, the system's response to a Taylor series cannot be reproduced by adding independently determined responses to each term.

Possibly the single most important advantage of the Taylor series description is that it provides for an ideal system having substance. The engineer soon realizes in treating any servomechanism, that it consists of an accumulation of physical limitations, e.g., limited actuator force, limited actuator speed, minimum actuator inertia, limited total power, etc. In Taylor series treatments, the ideal system has physical limitations identical with those of the actual system; it idealizes its control intelligence so that all operations are performed utilizing the system's maximum physical ability. For example, suppose the transient forcing function is a displacement step function requesting the output to move from its initial position a to some end position b. The system's best response is obviously to muster all available force (positive saturation), to accelerate as rapidly as possible, and then, precisely at the appropriate reversal point, to apply maximum braking force (negative saturation). With ideal intelligence, the output comes to rest at end point b. (For complex systems, the output ideally enters an end region about b and is forced to subside on point b by end-region controls. This end region is not considered here.) Contrasted with linear or extended linear treatments, this ideal accepts the physical handicaps of the

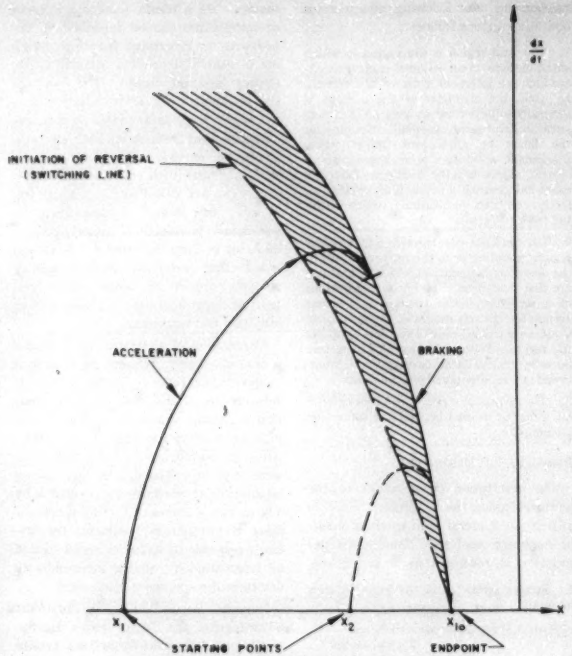


Fig. 1. Optimum hot-rod performance

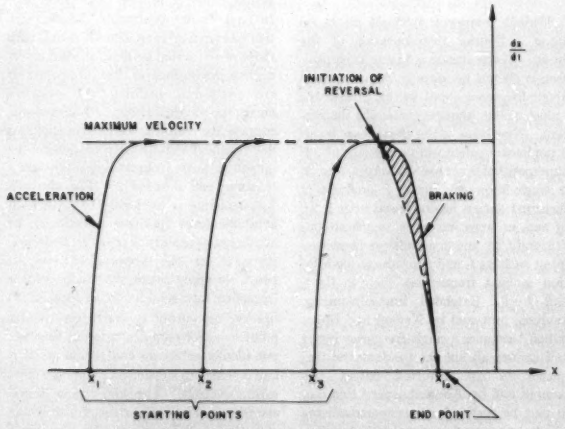


Fig. 2. Optimum elevator performance

existing servomechanism and utilizes the fact that these handicaps can be defined.

Practically all transient methods currently available fall into two categories: 1. those applicable to second-order systems having relay controllers, and 2. those applicable to higher order linear systems having relay or piecewise controllers. Phase-plane techniques, for example, are useful if the system can be described by a second-order quasi-linear (the second derivative or inertia term has a constant coefficient) differential equation. Analytic methods may be employed, usually in piecewise fashion, if the system is linear except for a switched-polarity constant-absolute-value input.

Description of the input by a power series is quite similar to description by a Taylor series. A section of the input function extending over time Δt is curve-fitted by a power series of time as the independent variable and the servomechanism is properly prepared by interpretation of the power-series terms. Although this approach has been suggested for linear treatments, very little has been mentioned in the literature concerning application to nonlinear problems.

Résumé

Before attempting an appraisal of the state of the art, two factors should be considered. Experience has shown, first, that the solution of a second-order linear servo problem is trivial in an engineering sense; good linear design considers fifth, tenth, and even higher orders. The effects of these higher order terms are not expected to lose their importance for nonlinear systems. Pruning down the description of a complicated phenomenon to a simple second-order form would likely introduce misleading results. Second, the descriptive adjective "linear" so pleasing to the analyst means trouble and expense to the designer and to the builder. When standard linear servomechanism techniques are used, it might be argued that these powerful tools justify the linear design constraint. However, discontinuously controlled and saturated systems do not require, nor are they necessarily improved, by linear components.

In summary, useful experimental frequency methods are available for treating certain nonlinear servomechanisms, but corresponding experimental transient methods have not been developed. Analytic and graphical methods which can be applied to transient studies of systems are restricted to second-order systems, certain third-order systems, and to linear systems having discontinuous or piece-

wise linear forcing functions. It is thereby argued that experimental transient methods would be useful; in the following, one is discussed.

An Experimental Transient Treatment of Saturated Responses

The postulated problem is to control a defined load or environment. The load may be linear or nonlinear and may have several degrees of freedom. "Controlled" means that one of its freedoms is measurable and is to be forced to specific positions or values requested by the input signal. It is further postulated that the actuator, the prime mover, has been selected and defined along with its coupling to the load. In this definition are included the various limitations of the actuator. Coupled to the actuator input is an amplifier (not necessarily electronic) which has features such as saturations and linear and nonlinear time lags and delays defined. However, here the unsaturated portion of the amplifier static gain characteristic is shaped. The amplifier accepts an error signal which is some function of the system static error (error between actual system output and requested system output) and other dynamic corrective functions. Both the static-error function and the dynamic functions are assumed to include unavoidable time lags, e.g., those inherent in an output-velocity-sensing element, but otherwise the functions are formable.

In short, the specified servomechanism dynamics may be equivalent to those described by a 10th- or 15th-order nonlinear equation. It should be emphasized that insensitivity of the method to order of the system should not be construed as a claim to include universally all single-loop nonlinear systems; many systems having special nonlinearities or undamped modes can be so devised that the approach would be poor or not applicable. However, for the servomechanisms investigated, it produced favorable results.

Throughout the entire procedure, the criterion of perfection is a practicable ideal system. This ideal system is identical with the actual system with respect to its many defined features, such as its load, actuator, saturations, time lags, etc. Its two controllable functions are undefined, but they are always assumed competent to direct the system to comply with transient input requests to the best of its physical ability. The only limitations of the ideal system are physical ones. It is hoped that some practicable "intelligence" in the actual system can be devised to provide the same responses.

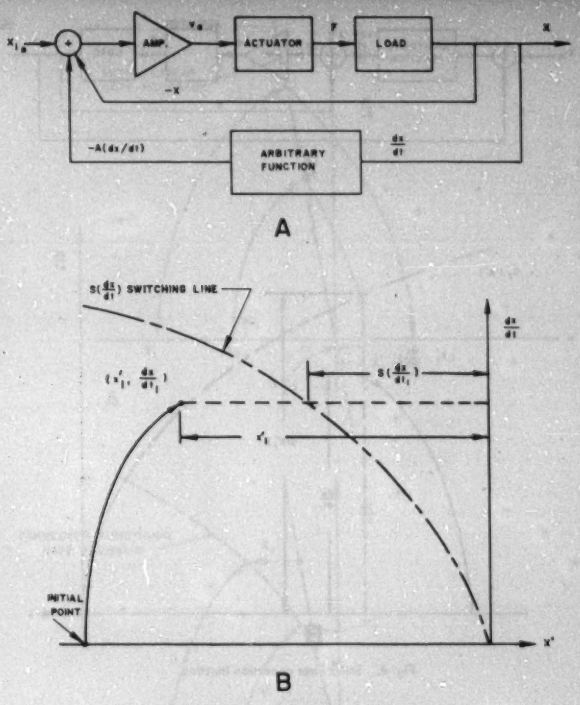


Fig. 3. Rate feedback corrective function

The Basis of the Method

The present method is thus concerned with attaining physically limited (or saturated) responses. It is a simple, 2-step procedure: 1. a set of system responses are recorded, and 2. from these records a "corrective" function is directly measured establishing the first reversal point. Near the null region, additional reversals are absorbed by allowing the unsaturated null region to be reasonably linear.

The method relies upon a work sheet which has the property of directly relating the system transient responses with intentionally added system corrective functions. A simple example is a system subjected to a displacement step (the first term of the Taylor series description). The work sheet is a system response plot of output velocity versus system static error. The mathematical implications of this plot as a phase-space cross section are ignored.

It is interesting to note that, faced with the simple problem posed in the foregoing, the skillful human operator approximates the ideal control. For example, consider the hot-rod driver who races for minimum transit time between stop signs a block or so apart. His automobile (actuator load) is capable of some maximum acceleration and some maximum braking deceleration. The system response on the work sheet, Fig. 1, shows that the skillful driver accelerates at full throttle to some precise reversal point and quickly transfers to full brake. The intelligence required is simply that needed to estimate the reversal point. If the block were shorter, the response and reversal point would change, as shown by the dotted line.

A second example is a competent elevator operator. His device accelerates rapidly to some velocity limit and then holds this limit until braking action is applied. Responses would be similar to

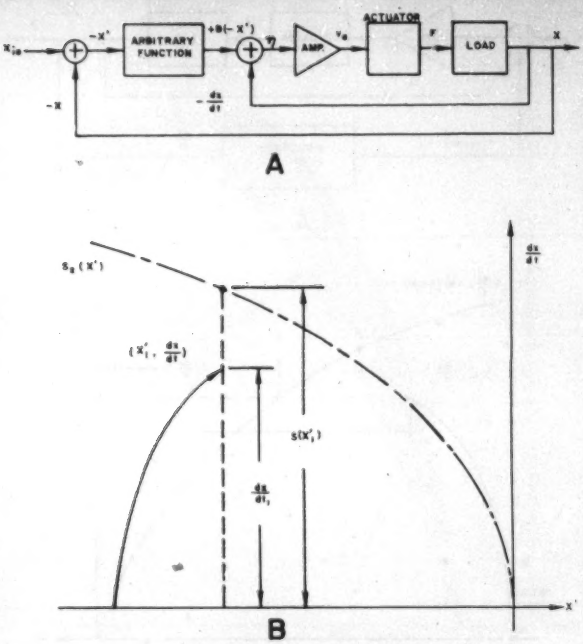


Fig. 4. Static error correction function

those shown in Fig. 2. Once maximum velocity is reached, the position of the reversal point relative to the end point (actually, end region) is fixed. Again the optimum intelligence simply selects this first reversal point.

If in the response plots $\dot{x}$ is plotted against system static error $x-x_{i_0}$, shifting the co-ordinates to the left, x_{i_0}, the end point is always the origin. Now responses to a variety of step inputs are plotted, each contributing one ideal switching point. The set of switching points will lie along a switching line, as shown in Figs. 1 and 2. A saturated accelerating trajectory (plot of a response) rising from any initial point x_i' that is switched to saturated braking when the trajectory crosses the switching line properly descends upon the origin. Therefore, it is concluded that the complete required intelligence is a switching line calling for a reversal of force when crossed by a trajectory. This switching line, as shown on the work sheet, can be thought of as either a function measured horizontally versus $\dot{x}$ the vertical, independent variable, i.e., $S_1(\dot{x})$, or as a function measured vertically versus x' as the independent variable, i.e., $S_2(x')$.

Corresponding to each of these interpretations of the switching function, there exists a simple physical means of providing the correction function. If the switching line is to be a function of output slot $S_1(\dot{x})$, the linear position feedback is selected plus some arbitrary function $A(\dot{x})$ of output velocity, as shown in Fig. 3(A). The error signal of this loop is

$$N = x_{i_0} - x - A(\dot{x}) = -x' - A(\dot{x}) \qquad (1)$$

Assume for the moment that the amplifier of Fig. 3(A) is a perfect relay which reverses its output v_a, thus forcing F when the error signal reverses polarity. In other words, the amplifier switches when $x' = -A(\dot{x})$.

The work sheet, Fig. 3(B), shows that a saturated accelerating trajectory must be switched when it crosses the ideal switching line, when $x' = S(\dot{x})$ if it is to be properly deflected to the end point. The control is obvious and very simple; the intentionally added rate feedback function $A(\dot{x})$ should be the negative of the ideal switching line $S_1(\dot{x})$. If a set of system response plots are made experimentally on the work sheet and the ideal switching line is drawn, this immediately and identically (except for sign) determines the appropriate rate feedback function.

If, on the other hand, the ideal switching line is chosen as a function of static error $S_2(x')$, the corrective function is an arbitrary function of static error, as shown in Fig. 4(A). In this case the error signal is

$$N = B(-x') - \dot{x} \qquad (2)$$

Again, when the amplifier is assumed to be a perfect relay, the force is reversed when N changes polarity, i.e., when $B(-x') = \dot{x}$. The work sheet, Fig. 4(B), indicates that the reversal should occur when $\dot{x} = S_2(x')$. Therefore, the intentional correction function of static error $B(-x')$ is exactly equal to the ideal switching function $S_2(x')$.

The point of these simplified examples is to show the direct relationship between the ideal switching line found by response tests and the intentionally added corrective function. The choice of using a functional rate feedback or a functional static error is optional with a designer. In actual systems, switching is not instantaneous for various reasons. If switching is effected by a relay, a certain amount of dead space and a certain transfer time exist. If switching is effected by a saturating amplifier, an unsaturated central region about the switch point exists. Time lags, delays, and velocity limits in the actuator further prolong the reversal time. The effect of these factors is a broadening of the instantaneous reversal line into a reversal band the finite width of which corresponds to a finite reversal time. However, the initiation of reversals is in all cases represented by a line, a series of decision instants, and therefore, for the experimental techniques discussed here, the ideal switching line can be considered as sponsoring the decision to reverse. All previous arguments hold.

In summary, if a system is subjected to a position step input, appropriate first switching information can be supplied to the system either by linear static error plus functional rate feedback, or by functional static error combined with linear rate feedback. The switching function should be formed so that all rising saturated trajectories are deflected directly towards the region of the origin, i.e., to the requested end point. This reasoning leads to a simple, direct experimental method of determining the best rate feedback function, or, alternatively, the best static error function.

An Experimental Method for Saturated Responses to Position Step Inputs

The complete servomechanism including linear rate feedback and linear static error is tested either as an analogue or as an actual system. Experimental velocities versus static error responses are recorded for various position step inputs.

In general, the recorded trajectories will not be perfect but will miss the origin (the requested end point) by certain static error increments $\Delta x_n'$ (error of the nth trajectory at zero velocity) dependent upon initial starting conditions, as shown in Fig. 5(A). If each of these positive velocity trajectory arches is shifted (here, to the left) by $\Delta x_n'$, the trajectories hit the origin; see Fig. 5(B). Such shifts in the switching points can be physically accomplished, as previously shown, by bending the rate feedback function. Reading off $\Delta x_n'$ misses immediately indicates the rate feedback correction ΔS_1 by the equality $\Delta S_1(\dot{x}_n) = -\Delta x_n'$, where $\dot{x}_n$ is the velocity at which the nth trajectory switches. A similar process would lead to the alternative static-error corrective function $S_2(x')$, but, for the sake of brevity, this approach is omitted; obviously, similar arguments and techniques are applicable to either.

Now consider tests of an actual system, a servomechanism which is reasonably linear for small inputs but is from time to time given input steps which drive its amplifier far into the saturated region. The particular system tested here included a linear inertia load, a 1-time-lag actuator, a 1-time-lag amplifier having fixed static saturation limits, and an output-velocity-sensing device having two time lags (slightly overcritically damped). (The amplifier time constant was 0.025 second. The actuator time constant was 0.005 second. Velocity-sensor time constants were 0.005 and 0.0025 second.)

The system is thus of the sixth order. Since here only the saturated region was of interest, the unsaturated linear region was treated by the usual procedure of determining maximum static gain such that, for small excursions, slightly underdamped responses occurred and further increase of rate feedback introduced second- (or velocity-) mode overshooting. With this prescribed static gain and linear rate feedback, a large step input sponsored a response with several characteristic overshoots, as shown in Fig. 6(A). For such a far-saturated response, the switching effect of the linear rate feedback is clearly displayed. A second response test was made which reduced the linear rate feedback by half and exaggerated the overshooting, as shown in Fig. 6(B).

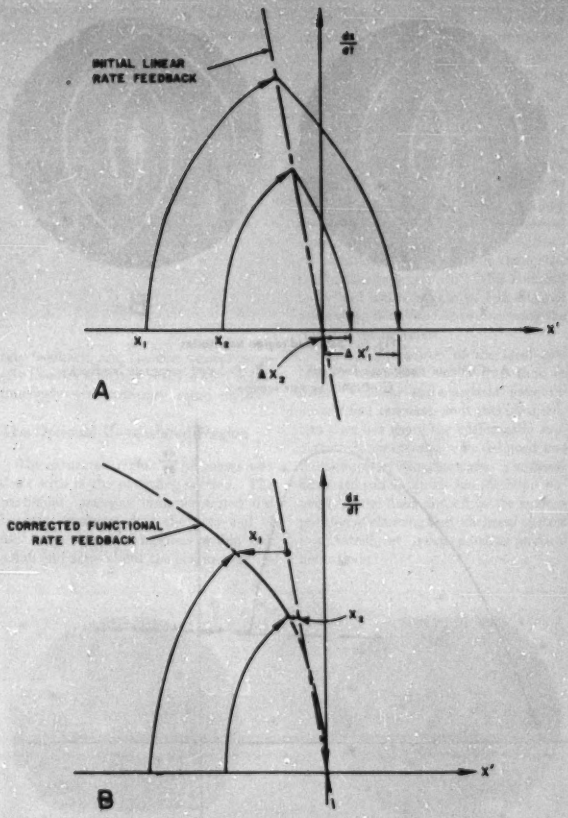


Fig. 5. Determination of functional rate feedback

These two simple tests gave nine points on the new nonlinear rate feedback function, enough to describe it accurately. The procedure was this: The response pattern of Fig. 6(A) was used as a starting point and each square of the grid was counted as a unit static error or a unit velocity; then the initial condition was at -8.8, 0. The trajectory was switched when its velocity was 9.9. The miss of this trial was measured when the trajectory arch again dropped to zero velocity $\Delta x_1' = 5.3$. This arch of the trajectory was now shifted intact 5.3 units to the left. The new starting point was $-(8.8 + 5.3) = 14.1$. As before, the new trajectory was switched when its velocity was 9.9, and it descended precisely on the origin. To produce this switching point, 5.3 units were added to the original rate feedback, i.e., to the linear feedback $a_1\dot{x}$. This addition gives one point of the new rate feedback function, i.e., when $\dot{x} = 9.9$, the amount of rate feedback should be $9.9a_1 + 5.3$. The second useful arch started where the first ended, at 5.3, 0. Its corresponding miss was 2.4 units, giving a second point on the feedback function at $\dot{x} = -7.7$. Nine usable arches occurred on the two figures providing the nine points of the feedback function, which described the function reasonably well, since, for this example, it was symmetric about the origin. Some new response trajectories are shown in Fig. 6(C).

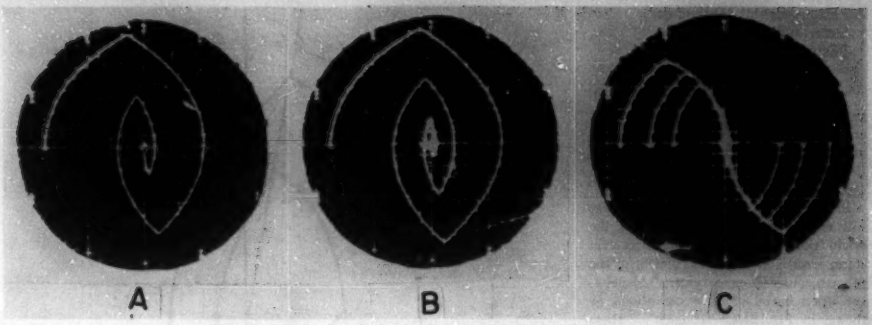


Fig. 6. Saturated region trajectories

A—Optimum linear rate feedback B—Reduced linear rate feedback
C—Nonlinear rate feedback

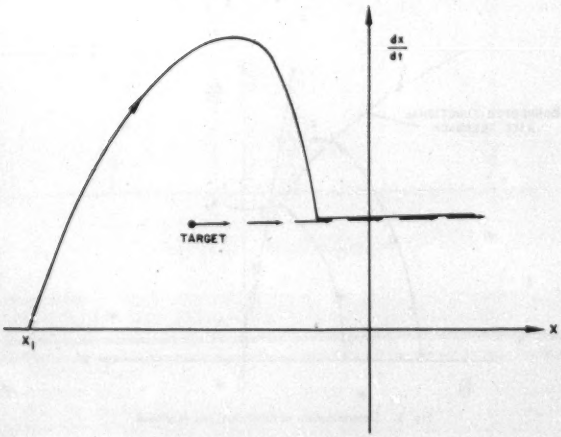


Fig. 7. Intercepting a constant velocity input

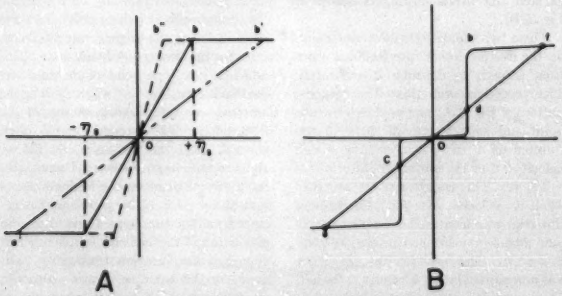


Fig. 8. Saturated linear and nonlinear static gain characteristics

Extension of the Method

The experimental procedure up to this point has been concerned with saturated systems subjected to position step inputs. Although not specifically stated, a further restriction is implied, namely, that the system load should not contain springs except as part of the couplings. If, for example, the load were an aerodynamic flipper subjected to nonlinear hinge moments (effectively a nonlinear torsion spring), the work sheets would have a different family of responses for each value of the end point. Consequently, there would be a different ideal switching function for each requested end point; in other words, the switching function would now be a 3-dimensional surface $S_1(\dot{x},x_4)$, rather than a line.

Similarly, each additional term of the Taylor series approximation of the input adds a dimension to the switching function. This addition is probably best illustrated by considering the next most simple representation of the input, i.e., the first two terms of the Taylor series.

The admission of the second Taylor-series term in approximating the input introduces the velocity step or ramp function. Unlike some of the published investigations of saturated systems responding to ramps alone, the problem here is that of treating systems subjected to an arbitrary displacement step input simultaneously with an arbitrary ramp input. If the work sheet co-ordinates, output velocity versus output position is used, the end point is not at the origin but is a target which at the beginning moves with some constant velocity, as shown in Fig. 7. The perfect system must intercept this target with saturated motions and only one reversal. The approach to this problem is straightforward.[1] A résumé of the conclusions here gives the more important results.

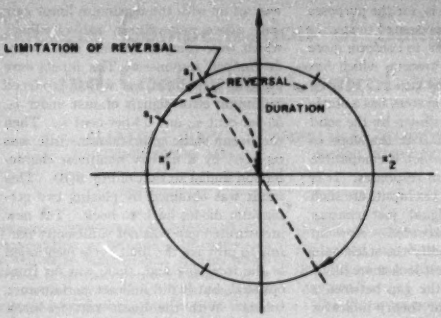


Fig. 9. Increased static gain limit cycle

If the system (amplifier-actuator-load) dynamics are subject to mathematical description, they can be described by $F(x^n, \ldots x^1, x)$ where $x^n = d^n x/dt^n$ and F denotes "some nonlinear function of . . ." The forcing function acting on this dynamic system is the error signal η. A simple argument[1] shows that the lowest order of x in the function F and the number of Taylor-series terms used to approximate the input determine the dimensional complexity of the switching function. This is shown in Table I. This

Table I. Number of Dimensions of Ideal Switching Function

Input Taylor Terms	x	$\dot{x}$	$\ddot{x}$	$\dddot{x}$
x_i	3	2	2	2
$x_i + \dot{x}_i t$	4	3	2	2
$x_i + \dot{x}_i t + 1/2 \ddot{x}_i t^2$	5	4	3	2

table brings out a curious feature: the switching function must have a minimum of two dimensions since it is not a constant. From the diagonal pattern of the table it could be inferred that 2 written under $\ddot{x}$ for a position step input is in some sense inferior to the 2 under, say, $\dddot{x}$ for a position step or the 2 under $\ddot{x}$ for a 2-term input. This is indeed so. Actually, it means that the switching function $S_1(\dot{x})$ determined for a $\ddot{x}$ type of system and a position step input are overdetermined. If the rate feedback function S_1 is used for this system, but with its independent variable changed to velocity error $\dot{\epsilon}$, where $\dot{\epsilon} = \dot{x} - \dot{x}_i$, a new ideal switching function is obtained which is suitable for step plus velocity inputs. For example, an analogue of a system having an inertial load and subjected to position step inputs has been described. For these tests, a switching function $S_1(\dot{x})$ was determined. Without further tests, this function may be employed as a function of velocityerr or $S_1(\dot{\epsilon})$ rather than of rate feedback, and the new system properly handles arbitrary position step simultaneously with arbitrary ramp inputs.

The Optional Unsaturated Region

The saturated region of responses was dealt with in the preceding section. The particular analogue example tested did become unsaturated at the very end of any trajectory, but motions within this small end zone about the origin were ignored. Now a magnified view of the unsaturated region is taken with the objective of determining first reversal points for small inputs. An engineer utilizing only linear techniques would confine system operation to this region and attempt to make the over-all system as linear as possible within this operating region. He would consider this the best practice. An engineer utilizing nonlinear transient techniques would attempt to eliminate the entire region since, within it, the system is not working at its maximum physical capabilities. He too would claim that this is the best practice. This confined battlefield is entered, not with the idea of attempting decisions but of discussing the rules for the decisions.

First, a comparison of the ideal systems is valueless. It is true that in theory a linear system could have infinitely fast responses, but this ideal system does not enjoy the enthusiastic support of engineers who have designed and built actual servomechanisms. The linear advocate cannot stress his ideal too fervently or he finds himself in the curious position of claiming that the linear system is so "intelligent" it can exceed its physical limitations.

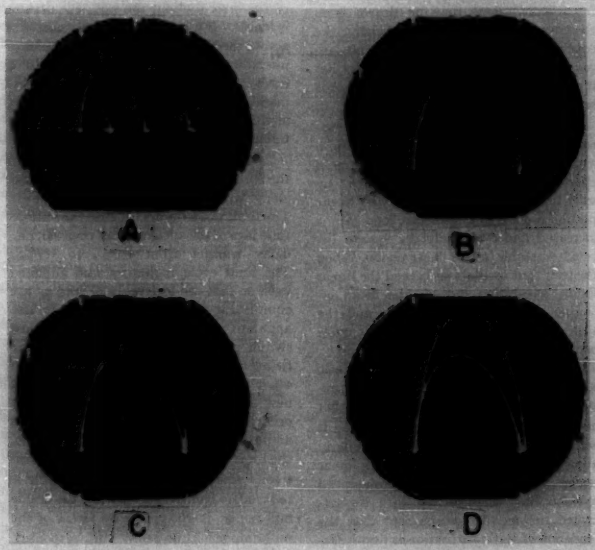


Fig. 10. Comparison of responses in the optional unsaturated region

A—Optimum linear feedback
B—Almost η_s input
C—50-per-cent η_s
D—25-per-cent η_s

Second, the system input must be defined. It can be argued but not proved that, in general, the system treated by nonlinear transient techniques can always have superior responses compared with a linearly treated system, provided that the inputs are appropriate step functions. This does not necessarily mean that the same superiority exists if the two systems are subjected to single-frequency inputs. Great care must be taken not to make generalizations based upon a few results. For example, some tests of specific systems, e.g., the system described in the following, may show improvements not only for transient responses but also for single-frequency input responses after being modified by transient nonlinear techniques. At the present state of the art, the explanation of this frequency response improvement and a definition of the class of systems for which this improvement occurs are both unclear. All that can be done is to test the various defined inputs in each system and compare the specific results.

Another comparison which must be handled with caution is that of conclusions reached by two different nonlinear approaches. Several conclusions may be drawn concerning system stability by utilizing extended linear analysis. Different conclusions may be drawn from the transient analysis. Superficially, the conclusions from the two methods appear to conflict; actually, the conflict lies in definitions of the problem and of words such as stability.

As a simple example, consider a servomechanism which is linear except for a saturating amplifier having a static gain characteristic, as shown in Fig. 8(A) by solid line. When the slope of the linear portion *a-o-b* is made steeper, loop gain is increased and eventually the system becomes unstable in this linear region. When the resultant oscillations exceed the linear limits, saturation reduces the mean effective loop gain of the system and the system settles at some stable limit cycle oscillation, as indicated by peaks a' and b'. Thus, it may be said that such static saturation is stabilizing in that it prevents divergent oscillations. On the other hand, in the earlier analogue example, it was observed that such a system properly adjusted for good responses in its linear region would respond by pronounced overshooting when driven into its saturated region. It can be said that the tendency of static saturation is destabilizing. The two entirely different operating conditions of the system and two interpretations of the word stabilizing produce conflicting conclusions.

The second example is, for the purposes herein, less trivial. It is desired to modify a basically linear system to conform more nearly with the ideal system, which has no optional unsaturated region. It is assumed that this linear system has a static gain characteristic, as shown by the solid line in Fig. 8(A), and that the slope of *a-o-b* is the maximum which is compatible with good linear system responses. It is further demanded that the inputs are such that η_s, the error signal just causing saturation, is never exceeded. Now an attempt is made to modify the static gain characteristic, to make it look more like a switch, by narrowing the gap between *a* and *b*. Extended linear theory indicates that narrowing this gap increases the effective gain between the new *a* and *b* limits and, eventually, will cause the system to become unstable in the unsaturated region. Instead of a new gain characteristic, such as that of Fig. 8(A), a shape such as that of Fig. 8(B) is chosen, which contains a very narrow unsaturated region *a-o-b* having a nearly flat (low-gain) null slope. When it is assumed that the mean effective gain between *a* and *b* is now considerably in excess of that causing system instability, it is observed that the system has two operating modes. For small inputs which sponsor only small excursions of the error η about the null, the system is stable. If a larger signal occurs driving across the barrier *c-d*, the system becomes unstable and, eventually, locks into a limit cycle peaking at *e-f*. Therefore, it might be concluded from extended linear analysis that such a static gain characteristic is undesirable.

In Fig. 9, the limit cycle mentioned in the foregoing is shown on the work sheet along with the linear switching line s_1. The limit cycle occurs because the various time lags of the system delay reversal, i.e., delay transfer from $+F_{max}$ to $-F_{max}$. The limit cycle trajectory had adjusted itself so that the intercepts x_1' and x_2' are equidistant from the origin. It is also apparent that there exists some new switch point s_1' which would direct the rising trajectory closer to the origin. In other words, by the proper selection of a new switching function, any limit cycle occurring in this system can be reduced in amplitude. By a series of such selections, the limit cycle can be reduced to any arbitrarily small value. If the limit cycle is reduced to confinement within the stable null region, it disappears by virtue of the well-damped (low-gain) null responses. As an experimental example, the previously mentioned sixth-order system used for saturation tests was also tested for unsaturated responses. First, the system was set up with the maximum linear gain and appropriate linear rate feedback which were compatible with slightly underdamped responses. The inputs were position step functions which produced maximum error signals of just under η_s, 50-per-cent η_s, and 15-per-cent η_s. Then the linear static gain characteristic was replaced by a narrow nonlinear characteristic similar to that of Fig. 8(B). This result was obtained by placing two germanium diodes back to back. The new unsaturated gap was not sufficiently narrow to provoke the limit cycle mentioned in the foregoing and, thus, was far from optional, but it did indicate performance trends. With the linear rate feedback dictated by linear treatment, response tests were made to the same step inputs as used for the linear tests; see Fig. 10(A). Since the saturated system moved more quickly than the linear, its responses tended to undershoot the end points.

Just as the far-saturated responses were treated by reading off the miss distance for each trajectory, the shift in x' required to bring the trajectories down on the end point was measured, and this measurement provided the correction to the original rate feedback. In contrast to the preceding saturated analysis, the correction here was negative, decreasing the rate feedback.

A comparison of responses to the displacement step inputs was made between the system incorporating nonlinear static gain, the newly determined rate feedback function, and the linear system optimized as previously described. New responses of both systems to the three different steps are shown in Figs. 10(B), 10(C), and 10(D). In all three instances, the inner trajectories were those of the linear system. Transit time is the integral of the inverse curve. Thus, the nonlinear system responses were in all cases more rapid than those of the linear system, particularly at smaller requested inputs where the latter was constrained to low forces.

References

1. Techniques in Treating Nonlinear Closed-Cycle Systems, R. S. Neiswander. *Ph. D. Thesis*, Parts I and IV, California Institute of Technology, Pasadena, Calif., 1954.

2. Introduction to Nonlinear Mechanics (book), Nicolas Minorsky. J. W. Edwards, Ann Arbor, Mich., 1947, chap. III.

3. Introduction to Nonlinear Mechanics (book), N. Kryloff, N. Bogoliuboff. Princeton University Press, Princeton, N. J., 1947, pp. 55–63.

4. A Frequency Response Method for Analysing and Synthesizing Contactor Servomechanisms, Ralph J. Kochenburger. *AIEE Transactions*, vol. 69, pt. I, 1950, pp. 270–84.

5. Optimization of Nonlinear Control Systems by Means of Nonlinear Feedbacks, R. S. Neiswander, R. H. MacNeal. *Ibid.*, vol. 72, pt. II, Sept. 1953, pp. 262–72.

Discrete Compensation of Sampled-Data and Continuous Control Systems

ELIAHU I. JURY
ASSOCIATE MEMBER AIEE

WILLIAM SCHROEDER
STUDENT MEMBER AIEE

Synopsis: This paper presents a method of compensation to be applied to continuous as well as to sampled-data systems which reduces to zero any overshoot or error after a prescribed finite time. This prescribed response is applicable to step or ramp inputs and could be extended to acceleration or higher order inputs. It is shown that this response can be accomplished using the modified z-transform[1] which yields the system response for all instants of time. The procedure, illustrated in examples, is applicable to general linear physical systems with no inherent delays. The compensated system may be designed so that shifts in gain or other dynamical disturbances have a small effect on the prescribed response. In conclusion, the method presented is powerful in synthesizing control systems to tally with a prescribed response and furthermore is straightforward in its application.

THE over-all objective of the paper is to present a useful and straightforward procedure for the synthesis of discrete compensation of fixed linear systems. The term "discrete" is chosen in order to include the use of either sampling compensators or continuous compensators consisting of a number of pure delay elements.

The problem of compensating a fixed system so that the step or ramp response meets predetermined specifications, namely rise time, overshoot, and settling time, is a familiar one. Transient response specifications are chosen, since the time response can be readily predicted and determined with the aid of modified z-transforms.[1] Fortunately, these specifications are common requirements of control systems. The importance of a straightforward and practical solution of such a problem needs no elaboration for the control system designer and should be recognized as very useful also to the designer of any other particular type of "system."

The literature contains a wealth of information for the compensation of continuous linear systems by the use of linear networks or devices. Nonlinear compensation is presently undergoing intensive research. Synthesis of inherent sampled-data systems and synthesis of continuous linear systems using digital or delay elements is relatively new. This paper is concerned with only the last two types of synthesis and is further restricted to a new method of determining how previously suggested components may be used to greater advantage.

Considering the previous work done on sampled-data systems, the most promising method of analysis appears to be the modified z-transform. Modification of the z-transform in order to include information at all times was first presented by Barker[2] and has been extended by one of the authors.[1] The desirable features of the modified z-transform as extended are that the response is defined at all times, and that the transient response can be expressed in the form of a contour integral in the z-plane. The evaluation of this integral can readily be predicted or obtained and the effects of certain alterations of the integrand can readily be seen without the evaluation of the integral. Thus a powerful tool has been made available for the synthesis of sampled-data and continuous systems. These features are not found in any other method and are the tools to be used in the procedure presented in this paper.

Analysis by the impulse response method,[3] or by other methods, may be more or less laborious, depending upon the particular system. However, they all have the disadvantage of being difficult to apply to synthesis. Analysis and synthesis with the z-transform, unmodified, is practical for some systems, but has the undesirable limitation of yielding the response at the sampling instants only. Bergen and Ragazzini[4] have jointly presented an interesting compensating procedure using a digital processing unit to obtain an over-all desirable transfer function. This procedure is, at best, applicable to systems whose time constants are greater than the sampling period and whose effective damping ratio is relatively large. By so restricting the original systems, the authors assume that the response between sampling instants will be well behaved. Once the design is completed, however, the system must be analyzed by some altogether different method to determine the validity of this assumption. Should the response between sampling instants be unacceptable, then no method is outlined for the improvement of the system. The interesting contribution of Bergen and Ragazzini is that an over-all desirable transfer function is selected as the objective, rather than any set of frequency response characteristics, which is essentially synthesis in the time domain.

A second method of compensation using a digital computer as the compensating element has been presented by Salzer.[5] However, it appears that the problem has been unnecessarily complicated by an attempt to apply conventional continuous system techniques and frequency response characteristics to sampled-data systems. The result of this appears to be that no better response can be obtained using the computer than could have been obtained by conventional continuous methods. A second limitation of Salzer's method is that experimental trial and error is required to determine the actual compensation to be used and to determine the response between sampling instants.

Considering previous work done on compensation of continuous systems by employing delay lines or other elements of pure delay, certain procedures and approaches have been presented by several different authors. However, as in sampled-data systems, all methods seem to have rather serious limitations. Recently Sze[6] has evaluated methods presented by other authors and has presented a graphical method of his own. Sze's objective is to obtain a close approximation of e^{-Ts} as the over-all transfer function since this is probably the next best thing to the normally physically impossible transfer function of one. Earlier authors in this field also selected this same objective, which, incidentally, is closely related with Bergen and Ragazzini's approach since the z-transform of e^{-Ts} is z^{-1}. The limitations of methods studied by Sze are relatively high gains required for fast responses, laborious analytical work for higher order systems, and the severe effect on the response of very small shifts in gain.

Paper 56-644, recommended by the AIEE Feedback Control Systems Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Summer and Pacific General Meeting, San Francisco, Calif., June 25–29, 1956. Manuscript submitted January 3, 1956; made available for printing April 17, 1956.

ELIAHU I. JURY is with the University of California, Berkeley, Calif., and WILLIAM SCHROEDER, formerly with the University of California, is now with the Ramo-Wooldridge Corporation, Los Angeles, Calif.

This research was supported by the United States Air Force, through the Air Force Office of Scientific Research, Air Research and Development Command, under Contract *AF18(600)-1521*.

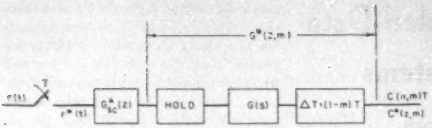


Fig. 1. Sampled-data system with series compensator and fictitious delay

Mathematical Background

It is generally known that the z-transform method offers a useful tool for the analysis of sampled-data systems, with the limitations of yielding information only at the sampling instants. As mentioned previously, the z-transforms can be modified[2] so as to yield the complete response of a sampled system by inserting a fictitious delay Δ in the forward path and, if feedback is used, a negative delay in the feedback path of the same amount. By employing operational calculus techniques, both the z-[3] and the modified z-transforms[1] have been extended and the following expressions for the modified z-transform have been developed. Referring to Figs. 1–4 and reviewing the modified z-transform,[1] one can write the modified transfer function in the form

$$G^*(z,\Delta)=\mathcal{L}[g(t-\Delta T)\delta_T(t)]$$
$$=\frac{1}{2\pi j}\int_{c-j\infty}^{c+j\infty}G(s)e^{-\Delta sT}\frac{1}{1-e^{sT}z^{-1}}ds \tag{1}$$

or let $m=1-\Delta$, then

$$G^*(z,m)=e^{-sT}\mathcal{L}[g(t+mT)\delta_T(t)] \tag{2}$$

$$G^*(z,m)=z^{-1}\frac{1}{2\pi j}\int_{c-j\infty}^{c+j\infty}G(s)e^{msT}\frac{1}{1-e^{sT}z^{-1}}ds \tag{3}$$

By real inversion[1]

$$C(n,m)T=\frac{1}{2\pi j}\int_{\Gamma}C^*(z,m)z^{n-1}dz \tag{4}$$

Where Γ is a contour which encloses all singularities of the integrand. For the series configuration shown in Fig. 1, the output-input relation is

$$C^*(z,m)=R^*(z)G_{sc}{}^*(z)G^*(z,m) \tag{5}$$

and for the over-all feedback system of Fig. 3,

$$C^*(z,m)=R^*(z)\frac{G_c{}^*(z)}{1+G_c{}^*(z)G^*(z)}G^*(z,m) \tag{6}$$

where $G^*(z)$ is $G^*(z,m)$ with m equal to unity.

Equations 5 and 6 are also applicable to Figs. 2 and 4 respectively, by replacing $G^*(z,m)$ with $G_0{}^*(z,m)$.

Equation 4 can be employed as a most useful tool. $C^*(z,m)$ can readily be obtained by substitution of equations 5 and 6 and the use of tables.[1,2] Furthermore, m is independent of the variable of integration, and can be treated as a constant in evaluating the integral. Thus the integral may be evaluated with tables,[7] long division,[8] or by the sum of the residue. Since m is a continuous variable within the range $0\leqslant m\leqslant 1$, the usual laws of differentiation may be applied. Thus if desired, the maximum response can be obtained either by equating $\delta C/\delta m$ to zero first and then evaluating the integral or the order of the operations may be reversed. Although evaluation of the integral is not generally a difficult task, other valuable information is obtainable from the integrand. For instance it can be shown that if all the terms of the integrand in Equation 4 which contain m have singularities at the origin only, then for all values of n above an obvious integer, the value of the integral is independent of m. This is a most useful feature. Also, the reader can determine the initial and final value of the response.[7]

Finally, since $0\leqslant m\leqslant 1$ and since n may take on an integer value, the complete response at any and all times can be found by evaluating equation 4 and the relation

$$t=(n-1+m)T \tag{7}$$

For these reasons the modified z-transform appears to be superior to any other tool for the synthesis of systems employing discrete compensators.

Development of Theory

A procedure for the compensation of sampled-data systems will be presented first. This same procedure is also applicable to continuous systems to be compensated with pure delay elements, as shown in a later section. The general configurations shown in Figs. 1 through 4 will be considered in this paper. It is realized that only the over-all feedback systems of Figs. 3 and 4 are of primary interest to the control system designer, however, examination of equations 5 and 6 reveals that they are of the same form. Determination of the feedback compensator by equation 6 requires first a determination of the quantity $G_c{}^*(z)/1+G_c{}^*(z)G^*(z)$ from which $G_c{}^*(z)$ may be determined.

It should be noted that the quantity $G_c{}^*(z)/1+G_c{}^*(z)G^*(z)$ corresponds to simply $G_{sc}{}^*(z)$ in equation 5. Thus the determination of the feedback compensator $G_c{}^*(z)$ yields as a by-product the series compensator $G_{sc}{}^*(z)$. For this reason, both a feedback and series system are designed at the same time, both of which will have the same response for a step input. This may be convenient if the series system merits consideration.

A straightforward approach then to synthesis with the configurations under consideration is first to determine the quantity $G_{sc}{}^*(z)$. Such a procedure is presented in this section, first for a step input and then for a ramp. The procedure could also be generalized to include acceleration and higher order inputs.

Step Response Synthesis

Assume that a fixed linear system, either open or closed loop, is available for some application but its step response must be improved in order to meet required transient specifications. Assume also that the input is available in either a continuous or periodically sampled form and the system's transfer function, $G(s)$, Figs. 1–4, is of the form:

$$G(s)=\frac{A_q s^q+A_{q-1}s^{q-1}+\ldots+A_0}{s^p(B_L s^L+B_{L-1}s^{L-1}+\ldots+B_0)} \tag{8}$$

where A and B are constants, $(p+L)\geqslant(q+1)$, s is the Laplacian variable, and all poles are in the left half of the s-plane, or at the origin (i.e., stable systems). The required step response is specified by (1) zero steady-state error, (2) maximum allowable rise time t_r, (3) maximum allowable overshoot M_p, and (4) maximum allowable settling time t_s. These quantities are illustrated in Fig. 5.

The general approach for synthesizing such a system will be to manipulate the integrand of the inverse transformation, equation 4, so that for all n larger than a fixed integer, the value of the contour integral is a constant, independent of n and m. This requires that all the terms of the integrand which contain m must have no contribution to the contour integration. The effect of this requirement is to cause the output to be flat or independent of time for all time greater than a fixed period.

A detailed procedure follows.

Step 1. If it is feasible to reduce the effective order of the original system by conventional methods, this will generally be desirable because it will result in a simpler compensator. This is not a necessary requirement and merely offers the designer additional freedom in obtaining the desired response.

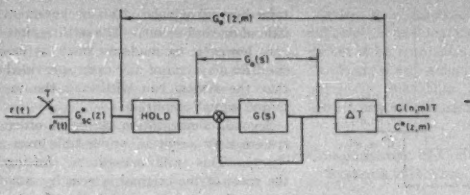


Fig. 2. Sampled-data system with series compensator and fictitious delay

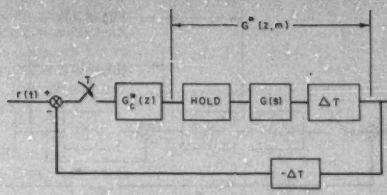


Fig. 3. Sampled-data control system with compensator $G_c^*(z)$ and fictitious delays

Step 2. Choose the period of the sampler to satisfy the relation

$$T=\frac{t_r}{(p+L)} \tag{9}$$

where t_r is the specified rise time and $(p+L)$ is as defined by equation 8. The justification for this choice is contained in equation 4 and will be further clarified in step 4.

Step 3. With the use of tables,[1] find $G^*(z,m)$, Fig. 1 or 3; the modified z-transform of $[G(s)\ (1-e^{-sT})/s]$. For systems whose $G(s)$ is of the form of equation 8, $G^*(z,m)$ will be of the form

$$G^*(z,m)=c\frac{z-1}{z}\left[\frac{1}{(z-1)^{p+1}}+\frac{\phi_p(m)}{(z-1)^p}+\cdots\frac{\phi_1(m)}{z-1}+\frac{\psi(m,z)}{(z-r_1)\cdots(z-r_L)}\right] \tag{10}$$

where p and L are as in equation 8, c is a constant, $\phi\ (m)$ and $\psi(m,z)$ are functions of m, and $r_1 \cdots r_L$ are necessarily in the unit circle of the z-plane since $G(s)$ is stable.[7]

Step 4. Choose $G_{sc}^*(z)$ so that: (1) all poles of terms of $G^*(z,m)$ which contain m are cancelled; (2) no new poles other than at the origin in the z-plane are introduced; (3) the steady-state error is zero; and (4) $G_{sc}^*(z)$ is physically realizable. If conditions 1 and 2 are met, then it can be seen from equation 4 that for all $n \geq (p+L+1)$ the response will be independent of both n and m (flat). This must be true since for all $n \geq (p+L+1)$, all terms containing m will have no singularities and the residue of $z^{n-1}/(z-1)z^{p+L}$ is unity. This is the essential feature of the procedure and is also the justification of step 2. The effect of conditions 1 and 2 may be more readily seen if the following general expression of $G_{sc}^*(z)$, which satisfies all four conditions, is examined in relation with equations 4 and 10.

$$G_{sc}^*(z)=\frac{1}{c}\frac{(z-1)^p(z-r_1)(z-r_2)\cdots(z-r_L)}{(1-r_1)(1-r_2)(1-r_3)\cdots(1-r_L)z^{(p+L)}} \tag{11}$$

The constant multiplier and the denominator are chosen so that the product of equations 10 and 11 is unity for z equal to unity. This satisfies the condition that the steady-state error is zero by the final value theorem.[7] (If evaluation of the inverse transformation by long division[8] rather than by contour integration is preferred, it can be seen that the coefficient of terms beyond $z^{-(p+L)}$ will be independent of m.) It is important to note that the poles of $G^*(z,m)$ are independent of m and can therefore be cancelled by $G_{sc}^*(z)$. Equally important is the fact that the zeros are functions of m and any attempt to cancel them will only result in the addition of undesirable poles. This is a limitation of the approach of Bergen and Ragazzini[4] which cannot be seen unless applying the modified z-transform. Condition 4 requires that the order of the numerator in z be no larger than the order of the denominator, which requires the introduction of poles at the origin.

By equations 5 and 6,

$$G_c^*(z)=\frac{G_{sc}^*(z)}{1-G_{sc}^*(z)G^*(z)} \tag{12}$$

from which $G_c^*(z)$ is realizable if $G_{sc}^*(z)$ is realizable.

Step 5. Obtain the complete step response by evaluating equation 4, as the following

$$[1/2\pi j\int_\Gamma z/z-1G_{sc}^*(z)G^*(z,m)z^{n-1}dz]$$

which will be very simple to do by the sum of the residues, by long division, or with tables since poles exist at the origin and at unity only. From the complete response the maximum can readily be determined and compared with the allowable M_p. An unacceptable overshoot presents the only possibility of the specifications being unsatisfied, since t_r and t_s will be satisfied. If the poles of $G(s)$ are all real and if the order of the denominator is at least two larger than the order of the numerator then the overshoot will be zero. If $G(s)$ contains complex poles, then there may or may not be an overshoot, depending upon the pole-zero location. If the overshoot is unacceptable, then it can be reduced by reducing the imaginary part of the most effective complex roots.

Step 6. After determining a $G_{sc}^*(z)$ which yields a satisfactory response, evaluate $G_c^*(z)$ from equation 12. This completes the procedure for determining the pulsed transfer function of the feedback compensator. The response of a series system with a compensator $G_{sc}^*(z)$ will be identical to the response of the feedback system, and vice versa, by equation 4.

Illustrative Example: To illustrate the simplicity and accuracy of the foregoing procedure a simple example will be considered. The compensation of higher order systems can be performed in the same manner, once $G(s)$ is expressed in a partial fraction expansion.

Assume the initial system is defined by a linear differential equation of second order, the transfer function of which is $G(s)=ca/s(s+2a)$ where c and a are both >0, and the desired step response is specified by $t_r=1$, $M_p=0$, and the steady-state error equal to zero. In general, t_r is arbitrary except $t_r>0$.

The desired response can be obtained by either leaving the poles of $G(s)$ as they are, or by shifting them prior to applying this method. If the poles are unchanged, then by employing step 2, T is chosen to be one half. From tables,[1] the modified z-transform is the equation

$$G^*(z,m)=\frac{c}{4}\frac{z-1}{z}\left[\frac{1}{(z-1)^2}+\frac{m-a^{-1}}{z-1}+\frac{e^{-2ma}a^{-1}}{z-e^{-2a}}\right] \tag{13}$$

and employing step 4, the transfer function $G_{sc}^*(z)$ is

$$G_{sc}^*(z)=\frac{4}{c}\frac{(z-1)(z-e^{-2a})}{(1-e^{-2a})z^2} \tag{14}$$

The step response is given by real inversion[1]

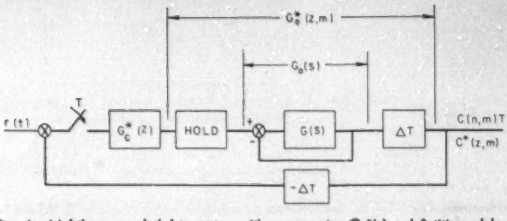


Fig. 4. Multiloop sampled-data system with compensator $G_c^*(z)$ and fictitious delays

$$C(n,m)T=\frac{1}{2\pi j}\times\int_\Gamma [R^*(z)G_{ac}{}^*(z)G^*(z,m)z^{n-1}]dz$$

$$=\frac{1}{2\pi j}\int_\Gamma\left[\frac{(z-1)(z-e^{-a})}{(1-e^{-a})z^2}\right]\times\left[\frac{1}{(z-1)^2}+\frac{m-a^{-1}}{z-1}+\frac{e^{-am}a^{-1}}{z-e^{-a}}\right]z^{n-1}\,dz \quad (15)$$

Evaluating[9] equation 15 for integer values of n,

$$n=0,\ C=0 \quad (16)$$

$$n=1,\ C=(m-a^{-1}+a^{-1}e^{-am})\frac{1}{1-e^{-a}} \quad (17)$$

$$n=2,\ C=1-e^{-a}(m-a^{-1})-a^{-1}e^{-am}\frac{1}{1-e^{-a}} \quad (18)$$

$$n\geqslant 3,\ C=1.0 \quad (19)$$

Note that for all $n>2$, ($t>1$, by equation 7, since the period is one half), the error is zero as specified. To determine the maximum response during the first and second periods, equate the partial derivatives of equations 17 and 18 to zero respectively,

$$n=1,\ \frac{\partial C}{\partial m}=0=1-e^{-am} \quad (20)$$

Solving for m from equation 20 yields a value of m equal to zero. For this value of m, the response is a minimum. The maximum response occurs at m equal unity, that is,

$$C_{max}=\frac{1}{1-e^{-a}}-\frac{1}{a} \quad (21)$$

From equation 21, C_{max} is less than unity for all a larger than zero. Similarly, during the second period,

$$n=2,\ \frac{\partial C}{\partial m}=0=-e^{-a}+e^{-am} \quad (22)$$

Solving for m from equation 22 yields a value of m equal to unity. This value of m yields the maximum response,

$$C_{max}=[1-e^{-a}(1-a^{-1})-e^{-a}-a^{-1}]\frac{1}{1-e^{-a}}=1.0 \quad (23)$$

Thus M_p equals zero for all positive values of a. The desired response can therefore be obtained with a compensator given by equations 12 and 14. Solving for $G_c^*(z)$,

$$G_c^*(z)=\frac{4}{c}\frac{z-e^{-a}}{(1-e^{-a})z+(a^{-1}-e^{-a}-a^{-1}e^{-a})} \quad (24)$$

The compensated system is shown in Fig. 6 and the step response of the compensated and uncompensated system is shown in Fig. 7. From equation 24, it is noticed that the compensator must store only one previous input and one previous output. As mentioned earlier in this part, the procedure also yields a series system with the same response, should this system be of interest. The series compensator is given by equation 14 and requires the storage of two previous inputs and no previous outputs.

The following merits and limitations of the procedure are recognized. First a flat response of zero overshoot after a prescribed finite time is obtainable. This cannot be done using linear compensation of similar systems and is physically difficult using nonlinear compensation.

A second advantage is that the designer has a large degree of freedom in obtaining the desired response. Only two basic configurations have been considered in this paper but there is no reason why the modified z-transform cannot be used equally well for more complicated configurations.[10]

Third, the response to any other input can be determined exactly, and it is possible to obtain a ramp response which has no ripple but has a constant steady-state error, as shown in Fig. 7.

The method also has limitations. First, since the rise time and settling time are independent of single amplitude, the fastest possible response is not obtained for all signals. This is characteristic of linear systems. The settling time can, however, be made as small as desired for a system of any order, provided that the system can withstand the required signal amplitude.

Second, compensation of higher order system may result in undesirable overshoots. This will necessitate shifting the roots of the original system by conventional methods and this may be difficult. This same problem would generally arise, however, if any other method of feedback compensation were used.

Ramp Response Synthesis

Assume the problem to be the same as for a step input except that the specified zero steady-state error, maximum allowable overshoot M_p, maximum allowable rise time t_r, and maximum settling time t_s are all with respect to a ramp input. These quantities are illustrated in Fig. 8.

In general, either series or feedback compensation may be used. The approach will be to manipulate the integrade of equation 4 so that for all n larger than a fixed integer, the value of the integral will be $T(n-1+m)$ for a unit ramp input. This requires in part that all the terms of the integrand which contain m, except the term $mT/z-1$, must have no contribution to the contour integration. Such a procedure is outlined in the following.

Step 1. Choose a first-order hold[9] circuit if the system does not have a pole at the origin in the s-plane. If the system does have a simple or multiple pole at the origin, then the designer is free to use either a zero order hold or remove the poles and use a first-order hold. The best choice will depend on the number of factors other than desired response.

Step 2. Choose the period of the sampler so that

$$T=\frac{t_r}{p+L+1} \quad (25)$$

where t_r is the specified rise time and $(p+L)$ is the order of $G(s)$.

Step 3. With tables, find $G^*(s,m)$, the modified z-transform of $G(s)\times 1-e^{-sT}/s$ or $G(s)\times[(1-e^{-sT})^2\cdot(A/s+1/Ts^2)]$, depending upon the order of the hold circuit.

Step 4. Choose $G_{ac}{}^*(s)$ so that: (1) all poles of terms of $G^*(s,m)$ that contain m are cancelled if a zero order hold is used and all except a first-order pole at unity are cancelled if a first order hold is used; (2) no new poles other than at

the origin are introduced; (3) one adjustable constant exists or is introduced so that the steady-state error can be made zero for all ramps; $G_{ss}^*(z)$ is physically realizable. Condition 1 and 2 will cause the response to have the same steady-state rate as the input but with a constant lag, depending on the system. As with the step input this can be seen by examining equation 4. In order to make the error be zero, some arbitrary constant must be introduced either in the compensator or the hold circuit. If a zero order hold is used, then the constant may be introduced as an additional zero in $G_{ss}^*(z)$ and if a first-order hold is used it may be introduced in the hold circuit.

Step 5. Obtain the complete response and evaluate the arbitrary constant for zero error. Determine the maximum overshoot, and if unacceptable, again it can systematically be reduced (to zero if necessary) by decreasing the imaginary part of the effective complex roots.

Step 6. Solve for $G_c^*(z)$ from equation 12.

An illustrative example is given in the Appendix.

The procedure for obtaining a specified ramp response has the same merits as the step response procedure with the addition of being able to obtain zero error, zero overshoot, and finite settling time for a ramp input.

However, the procedure also has additional limitations.

The first is that a system without an integrator must receive an unlimited signal if it is to follow a ramp indefinitely. This may be acceptable if the ramp inputs are restricted in time and amplitude so that the system will not reach saturation or if the original system is itself a closed loop system and the forward transfer function has a pole at the origin. This limitation is also true of any other method of compensation and it should be noted that only a first-order integrator is required to avoid this difficulty.

The second limitation is that both zero error and zero overshoot cannot be obtained simultaneously for step and ramp inputs. The ramp procedure automatically insures that the system will have zero steady-state error and the same settling time for both step and ramp inputs. However, the transient overshoot to a step input may be as much as 100 per cent for a system which has zero ramp overshoot. The step overshoot can be made as small as desired (but not zero) by decreasing the period and adding additional zeros to the compensator.[9]

A less serious limitation might be the difficulty in realizing the first-order hold, should it be required. Although not as simple as the zero order hold, it is certainly physically realizable and should not present any serious problem.

Compensation of Continuous Systems with Pure Delay Elements

The step and ramp procedures presented in the preceding sections are applicable to sampled-data systems. This section of the paper will be concerned with the use of these same procedures (with limitations) to the synthesis of continuous systems compensated with pure delay elements.

First consider Figs. 9 and 1. The two systems will have exactly the same response to a step input if the period of the sampler is Δ_1, $G_{ss}^*(z)$ is the z-transform of the pure delay compensator, and if the hold circuit is a zero order. This must be true since the input to $G(s)$ will be exactly the same for both systems. Likewise the responses of the two systems to a ramp input will be identical if the hold circuit accurately converts a sampled ramp into a continuous ramp. The transfer function of such a circuit is $e^{sT}(1-e^{-sT})^2/Ts^2$ which can easily be deduced from the Laplace transform of a sampled ramp[9] $Te^{-sT}/(1-e^{-sT})^2$. The fact that such a hold circuit may not be physically realizable is of no consequence, since the sampled system shown in Fig. 1 is to be used only as a mathematical equivalence of the continuous system. It should be noted that the sampled system is equivalent for a particular input only, once the hold circuit has been specified. For any other output, a different hold must be considered in order for the systems to be equivalent.

Therefore, if the problem is to compensate a fixed system so that the step or ramp response will meet prescribed specifications, as assumed in the previous sections, then the designer can use the procedures presented for determining $G_{ss}^*(z)$, with two exceptions. First, a zero order hold must be used for a prescribed step response and a hold circuit whose transfer function is $e^{sT}(1-e^{-sT})^2/Ts^2$ must be used if a specified ramp response is desired. Second, only series compensation can be used, since a hold circuit which reproduces the sampled error exactly is difficult to find.

After $G_{ss}^*(z)$ has been determined, the configuration of delay elements can be obtained simply by substituting e^{sT} for z.

Illustrative Example: Consider the same problem and system as used to illustrate the step response synthesis, except that the input is available continuously. Since only series compensation is to be considered, the original system will first be made a closed loop system. If the gain is adjusted so that the imaginary part of the complex roots is π/T ($T=1$), a sampled system whose

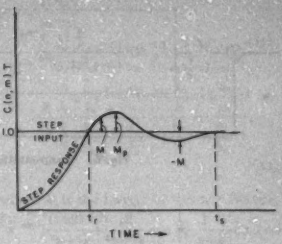


Fig. 5. Step response quantities

$$G_{ss}^*(z)=\frac{z+e^{-a}}{(1+e^{-a})z} \tag{26}$$

will yield the desired response. To obtain the desired equivalent continuous system, simply remove the hold circuit and sampler and replace $G_{ss}^*(z)$ with $G(e^{sT})$. Therefore,

$$G(e^{sT})=\frac{1+e^{-a}e^{-s}}{1+e^{-a}} \tag{27}$$

This continuous system, shown in Fig. 10, will have the same step response as a sampled system with a $G_{ss}^*(z)$ given by equation 26 and a zero order hold. The ramp response can be obtained by determining the ramp response of a sampled system with the same $G_{ss}^*(z)$, but with a hold circuit whose transfer function is $e^{sT}(1-e^{-sT})/Ts^2$. Such an equivalent system is shown in Fig. 11. To obtain the ramp response, first obtain $G_0^*(z,m)$ as follows:

$$G_0^*(z,m)=\frac{(z-1)^2}{z}\left[\frac{1}{(z-1)^2}+\frac{m-\dfrac{2a}{a^2+\pi^2}}{z-1}+\frac{\dfrac{2a}{a^2+\pi^2}e^{-am}\dfrac{\cos m\pi-\dfrac{a^2-\pi^2}{2a\pi}\sin m\pi}{z+e^{-a}}}{}\right] \tag{28}$$

Then by equation 4,

$$C(n,m)T=\frac{1}{2\pi j}\int_\Gamma R^*(s)G_{ss}^*(s)G_0^*(z,m)z^{n-1}dz \tag{29}$$

Evaluating for integer values for n

$$n=0,\ C=0 \tag{30}$$

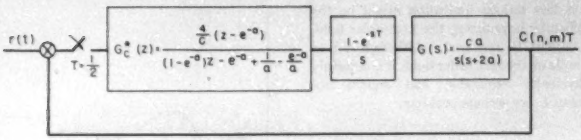


Fig. 6. Compensated second-order system

$$n=1,\ C=\frac{1}{1+e^{-a}}\left[m-\frac{2a}{\pi^2+a^2}+\frac{2a}{\pi^2+a^2}e^{-am}\times\left(\cos m\pi-\frac{\pi^2-a^2}{2a\pi}\sin m\pi\right)\right] \quad (31)$$

$$n\geq 2,\ C=\frac{1}{1+e^{-a}}+m-\frac{2a}{\pi^2+a^2}+n-2 \quad (32)$$

This response, with the step response, is shown in Fig. 12.

Compensation of continuous systems with real delay elements by the above procedure has several desirable features, in addition to the advantages of the procedures for sampled systems. One of these is simplicity of realization. No hold or sampler is required and the compensation should not be difficult for electromechanical systems. Tape recorders or other types of magnetic storage are possible and advantage could be taken of the current research for real delay elements. A second advantage is that step response of a system which has been synthesized for a zero error ramp response is superior to the step response of a sampled system.

The main disadvantage of the procedure is that it is not applicable to the synthesis of over-all feedback systems. Series compensation requires that the compensator must be as precise as the desired response, and that any integration in the original system must first be removed. An approximate method for the synthesis of continuous over-all feedback systems has been devised and will be presented in a later report.

If the series procedure is compared with Sze's method[8] of series compensation, the outstanding advantage of the foregoing procedure is that the compensated system can be made much less critical to shifts in the dynamics of the original system or shifts in gain. Sze obtains a fast rise time and a flat response thereafter by applying large positive signals followed by comparable negative signals. The response of such a system will, in general, be more sensitive to shifts in roots than will a system whose input is always positive. This is discussed in the following section.

Discussion of Assumptions

The transfer function of the compensator is chosen so that certain undesirable poles of the original system are cancelled. In any physical system, it is realized that these poles will never be cancelled exactly, and therefore an investigation of the effect of imperfect cancellation should be made for each compensated system under consideration. Again, the modified z-transform is a useful tool since the complete response can be found for any pole-zero configuration.

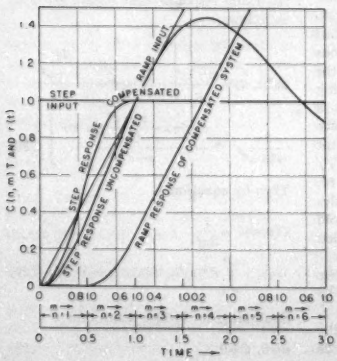


Fig. 7. Response of compensated second-order system shown in Fig. 6

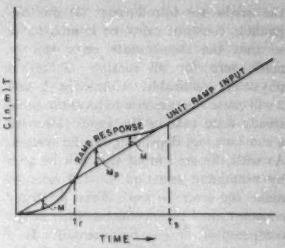


Fig. 8. Ramp response quantities

As pointed out in references 11 and 12, imperfect cancellation of a pole either outside, on, or near the unit circle in the z-plane can result in an unstable system. Therefore the compensator should never be required exactly to cancel a pole in this region of the z-plane. If the original system is stable, as assumed in the section on development of theory, then the modified z-transform can have no poles outside the unit circle. However, if the original system has a multiple pole at the origin in the s-plane, $G^*(z,m)$ will have multiple poles at $1+j0$ in the z-plane. If the procedure of the afore-mentioned section for a step input is applied directly, at least one of these poles will require cancellation. Cancellation should not be attempted however, since imperfect cancellation can result in an unstable system. This difficulty can be circumvented in either of the following ways and possibly others.

1. The order of integration of the original system may be reduced to one or zero with a continuous feedback loop. The resulting system must be stable, otherwise there will be a pole outside the unit circle requiring cancellation. A single order of integration gives no difficulty for a zero error step response, as noticed from the first example and equation 25.
2. The multiple integration can be retained and the system compensated for zero error for a ramp or higher order input so that the multiple integration is utilized. This will result in an overshoot for step inputs. The step overshoot can be reduced by adding additional zeros to the compensator but this increases the complexity of the compensator.

Imperfect cancellation of poles inside the unit circle in the z-plane can cause undesirable responses even though the system remains stable. This is particularly true of series compensation, where imperfect cancellation results in both a transient and steady-state error. The error or effect of the noncancellation will be considerably less if the output of the compensator is a series of positive steps or

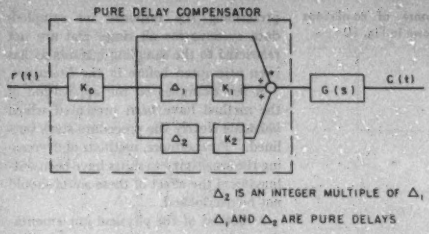


Fig. 9. Continuous system compensated with pure delay elements

$$G_c^*=\frac{C_c^*(z)}{R_c^*(z)}=\frac{a_0+a_1z^{-1}+a_2z^{-2}+a_3z^{-3}+a_4z^{-4}}{b_0+b_1z^{-1}+b_2z^{-2}+b_3z^{-3}+b_4z^{-4}}=\frac{\frac{a_0}{k_1}+\frac{a_1}{k_1}z^{-1}+\frac{a_2}{k_1}z^{-2}+\frac{a_3}{k_1}z^{-3}+\frac{a_4}{k_1}z^{-4}}{\frac{k_1}{k_2}\left(1+\frac{b_1}{b_0}z^{-1}+\frac{b_2}{b_0}z^{-2}+\frac{b_3}{b_0}z^{-3}+\frac{b_4}{b_0}z^{-4}\right)} \quad (33)$$

where k_1 is the largest of the a_i. Solving for $C_c^*(z)$ yields

$$C_c^*(z)=\frac{k_1}{b_0}\left(\frac{a_0}{k_1}+\frac{a_1}{k_1}z^{-1}+\frac{a_2}{k_1}z^{-2}+\frac{a_3}{k_1}z^{-3}+\frac{a_4}{k_1}z^{-4}\right)-\frac{k_2}{b_0}\left(\frac{b_1}{k_2}z^{-1}+\frac{b_2}{k_2}z^{-2}+\frac{b_3}{k_2}z^{-3}+\frac{b_4}{k_2}z^{-4}\right)C_c^*(z) \quad (34)$$

where k_2 is the largest of the b_i. By factoring out the largest coefficients k_1 and k_2, all others are less than unity and therefore only two amplifiers will be required. One configuration is then as shown in Fig. 13. The sampler may be placed on either end. If it is in the input, the delay must be nearly exact integers of the sampling period T since the pulses must be synchronized upon arrival at the summing devices. Likewise the amplifiers may be altered in position, depending upon the desired summing level. The approach is not restricted to electric systems, and the delays may, of course, be realized in either a digital or analogue method or combination of both.

A second approach is to consider the realization as a computing problem since the z-transfer function can be considered a computing program. A digital computer with a set program could therefore be used, but this may or may not be economically feasible. An alternate solution would be the design of a computer, either digital or analogue, which has the desired fixed program. The essential feature of such a compensator is that it must

impulses,[8] rather than both positive and negative, for a step input. This would be expected for a system of any order, since the algebraic sum of both positive and negative quantities will be more affected by a shift in one of the quantities than would the sum of only positive quantities. A series system can therefore be made less sensitive to shifts in the system by using a compensator whose output is always positive for a step input.

Over-all feedback systems have the advantage of being far less sensitive to imperfect cancellation than open-loop systems. So long as the imperfect cancellation does not cause instability, the steady-state error for step input will be unaffected. If the system as originally designed is found to be too sensitive to shifts or drifts resulting in imperfect cancellation, it can generally be made less sensitive by the addition of zeros in the transfer function of the compensating device and an increase in the sampling rate.[9]

In both the step and ramp response synthesis, it was tacitly assumed that the period of the sampler could be varied. If the period cannot be changed and the desired response cannot be obtained with the fixed period, it may be possible to reduce the effective order of the system by adding a minor feedback loop and considering only the dominant poles.

It has also been assumed that the system is linear. Reduction of the rise time by reducing the sampling period will require higher gains and therefore saturation will result with smaller signals. The effect of saturation has not been investigated but saturation and the maximum forcing functions which the system components can withstand will limit the rise times obtainable.

Realization of the Discrete Compensator

The preceding sections have been concerned primarily with the determination of the transfer function of the compensator in order to obtain a desirable response. With the exception of converting the transfer function into a continuous configuration of pure delay elements, no mention has been made of the physical realization of the compensator. This problem can be examined in detail only when a specific system is under consideration, therefore only broad general considerations will be discussed in this paper.

The realization of the required pulse transfer function offers considerable freedom, even if only electric systems are considered. There appears to be at least three different approaches to the problem and each approach offers a large degree of freedom.

First, any transfer function whose numerator and denominator is a rational polinomial in z, with the denominator of equal or higher order than the numerator, can be realized with pure delay elements, voltage dividers, two algebraic summing devices, and two amplifiers. (If the sampler and hold are considered as a part of the compensator, then these elements will also be required. If the system is originally a sampled one, then these elements will be required or present even if the system is not compensated.) The arrangement of configuration and the value of the elements is apparent if the output of the compensator is equated to the algebraic sum of previous outputs and inputs. As an example, consider the following:

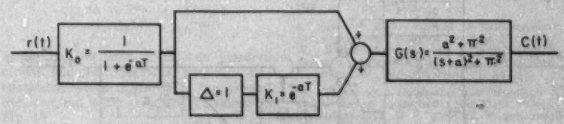


Fig. 10. Compensated continuous system

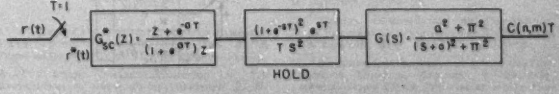


Fig. 11. Sampled-data equivalent of Fig. 10 for a ramp input

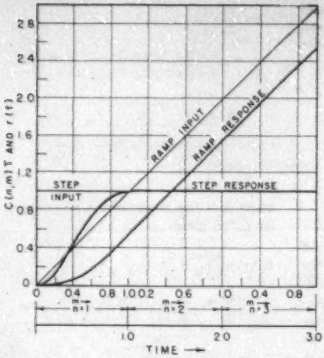


Fig. 12. Response of continuous system shown in Fig. 10

contain some facility for the memory of storage of previous quantities and must be able to handle or process this information so that the output is the sum of properly weighted quantities. Since the computing field is rapidly expanding, with an increasing number of devices for storing and handling information, the designer enjoys wide freedom in actual physical realization of the compensator.

A third approach to the realization is to make use of the existing tables of z-transforms and thereby obtain a corresponding Laplace transfer function. By synthesizing a network which has the desired Laplace transfer function, the desired z-transfer function can be realized by sampling both the input and the output of the network with synchronized switches. It may or may not be possible to obtain a corresponding Laplace transform which is physically realizable with a passive network. Since a large number of Laplace transforms can be found which have the same z-transform and since Laplace transfer function can generally be realized by more than one network, again this approach offers considerable freedom. In fact, Sklansky[13] has presented a detailed method whereby any physically realizable z transfer function may be realized by a combination of resistance-capacitance networks, zero order holds, summing or subtracting devices, constant gain amplifiers, and possibly pure delay elements.

Conclusions

The method of compensation presented in this report is a powerful tool in synthesizing continuous systems, or sampled-data feedback systems to yield a prescribed response. The prescribed response of zero overshoot and finite settling time is of much importance in the application of feedback systems. The procedures discussed in this report are straightforward and easy to apply for compensating control systems to yield such responses. They have the advantage of prescribing the response for sampled-data systems for all times and are not restricted to the sampling instants as has been attempted before in the literature.

A few examples on the application of the method have been presented which illustrate clearly the procedure steps outlined. Furthermore, methods of decreasing the sensitivity to shifts have been outlined[3] and the effect of these shifts should not be overlooked.

A method of the physical implementation of the compensators is presented which indicates the ease with which these compensators can be actually constructed.

The same procedure can be extended to include the compensation of systems with inherent delays in the forward or feedback paths which are not necessarily an integer value of the sampling period. Furthermore, as mentioned in the section on compensation of continuous systems, an approximate method can be devised whereby continuous systems may be compensated with pure delays in the forward path of closed loop systems.

The problems which have yet to be explored are a compromise between rise time and overshoot by cancelling only part of the poles, the effect of noise in the input, compensation of nonlinear systems, compensation by employing a variable pulse width, approximate realization of the compensator with continuous networks, and application of this method to the effective broadbanding of communication networks and systems without serious loss of gain.

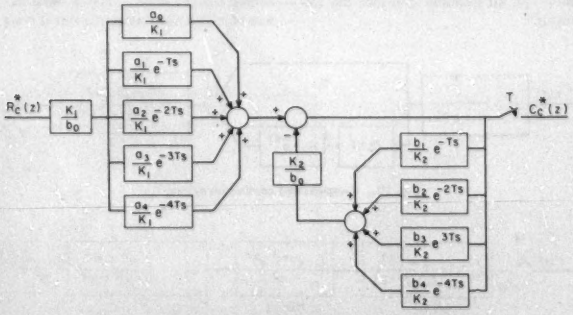


Fig. 13. Realization of discrete compensator with pure delays

Appendix

To illustrate the procedure for obtaining a desired ramp response, consider the same system as used to illustrate the step response procedure: $G(s)=ca/s(s+2a)$, k_r is unity, c, a, and k_r are arbitrary except greater than zero. Since the system has a pole at the origin, a zero order hold and a period of one third are chosen (steps 1 and 2). By tables,[1]

$$G^*(z,m)=\frac{c}{6}\frac{z-1}{1}\left[\frac{1}{(z-1^3)}+\frac{z^{m-\frac{3}{2a}}}{z-1}+\frac{3e^{-2am/3}}{2a(z-e^{-2a/3})}\right] \quad (35)$$

From step 4,

$$G_{sc}^*(z)=\frac{6}{c}\frac{(z-1)(z+A)(z-e^{-2a/3})}{(1-e^{-2a/3})(1+A)z^3} \quad (36)$$

The ramp response is given by equation 4,

$$C(n,m)T=\frac{1}{2\pi j}\int_\Gamma R^*(s)G_{sc}^*(z)G^*(z,m)z^{n-1}dz$$

$$=\frac{1}{2\pi j}\int_\Gamma\frac{(z-e^{-2a/3})(z+A)}{(1-e^{-2a/3})(1+A)3}\times$$

$$\left[\frac{1}{(z-1)^2}+\frac{m-\frac{3}{2a}}{z-1}+\frac{3e^{-2a/3}}{(z-e^{-2a/3})}\right]z^{n-1}dz \quad (37)$$

Evaluating the output for integer values of n yields for

$$n=0, 1;\ C=0 \quad (38)$$

$$n=2;\ C=\frac{1}{(1-e^{-2a/3})(1+A)}\times\left[\frac{m}{3}-\frac{1}{2a}+\frac{e^{-2am/3}}{2a}\right] \quad (39)$$

$$n=3;\ C=\frac{1}{(1-e^{-2a/3})(1+A)}\times\left[\frac{1}{3}+(1+A-e^{-2a/3})\left(\frac{m}{3}-\frac{1}{2a}\right)+\frac{A}{2a}e^{-2am/3}\right] \quad (40)$$

$$n\geqslant 4, C=\frac{(2+A-e^{-2a/3})}{3(1-e^{-2a/3})(1+A)}+\frac{m}{3}-\frac{1}{2a}+\frac{1}{3}(n-4) \quad (41)$$

For zero error, the output must equal the input, that is

$$C=\frac{1}{3}(n-1+m) \quad (42)$$

Equations 41 and 42 yield

$$\frac{2+A-e^{-2a/3}}{3(1-e^{-2a/3})(1+A)}-\frac{1}{2a}-1=0 \quad (43)$$

and solving for A from equation 43 gives

$$A=\frac{1-e^{-2a/3}}{3(1-e^{-2a/3})\left(1+\frac{1}{2a}\right)-1}-1 \quad (44)$$

It is noticed from equation 44 that for all a greater than zero,

$$-1<A\leqslant-\frac{1}{2} \quad (45)$$

To determine the overshoot M during the second and third periods, respectively,

$$n=2,\ M=C-\frac{1}{3}(1+m)=\frac{m-\frac{3}{2a}+\frac{3}{2a}e^{-2am/3}}{3(1-e^{-2a/3})(1+A)}-\frac{1}{3}(m+1) \quad (46)$$

To obtain the maximum overshoot, equate

$$\frac{\partial M}{\partial m}=0 \quad (47)$$

which yields M_{max} less than zero for the second period. Similarly, during the third period,

$$n=3,\ M=\frac{1+(1+A-e^{-2a/3})\left(m-\frac{3}{2a}\right)+\frac{3A}{2a}e^{-2am/3}}{3(1-e^{-2a/3})(1+A)}-\frac{1}{3}(m+2) \quad (48)$$

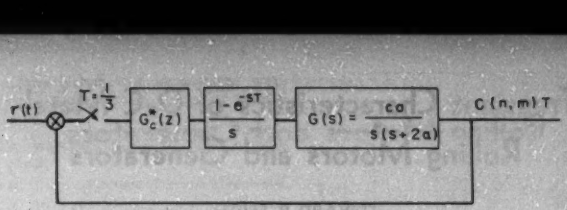
and solving for the maximum overshoot,

$$\frac{\partial M}{\partial m}=0 \quad (49)$$

yields m equal unity for which M_p is zero. Thus the desired response can be obtained with a $G_{ss}{}^*(s)$ given by equations 36 and 44. The transfer function of the compensator can be obtained with equation 12,

$$G_c{}^*(s)=\frac{\frac{12a}{c}(s-e^{-2a/3})(s+A)}{2a(1-e^{-2a/3})(1+A)s^2+[2a(a-e^{-2a/3}-Ae^{-2a/3})]s+A(3-3e^{-2a/3}-2ae^{-2a/3})} \quad (50)$$

which requires storage of two previous inputs and two previous outputs. The completed system is shown in Fig. 14 and the step and ramp response are shown in Fig. 15.



$$G_c^*(z)=\frac{\frac{12a}{c}(z-e^{-2aT})(z+A)}{2a(1-e^{-2aT})(1+A)z^2+\left[2a(A-e^{-2aT}-Ae^{-2aT})\right]z+A(3-3e^{-2aT}-2ae^{-2aT})}$$

$$A=\frac{1-e^{-2aT}}{3(1-e^{-2aT})(1+\frac{1}{2a})-1}-1$$

Fig. 14. Compensated second-order system

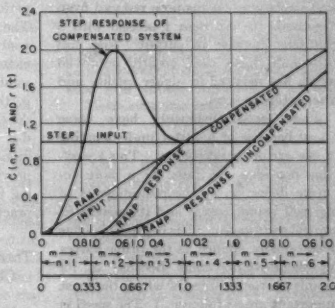


Fig. 15. Response of compensated second-order system shown in Fig. 14

References

1. Synthesis and Critical Study of Sampled-Data Control Systems, Eliahu I. Jury. *AIEE Transactions*, vol. 75, pt. II, July 1956, pp. 141–51.

2. The Pulse Transfer Function and Its Application to Sampling Servosystems, R. H. Barker. *Proceedings*, Institution of Electrical Engineers, London, England, vol. 99, pt. IV, 1952.

3. The Design of Sampled-Data Feedback Systems, Gladwyn V. Lago, John G. Truxal. *AIEE Transactions*, vol. 73, pt. II, Nov. 1954, pp. 247–53.

4. Sampled-Data Processing Techniques for Feedback Control Systems, Arthur R. Bergen, John R. Ragazzini. *Ibid.*, pp. 236–47.

5. System Compensation with a Digital Computer, J. M. Salzer. *IRE Convention Record*, Institute of Radio Engineers, New York, N. Y., pt. 5, pp. 179–86.

6. Delay Line Compensation in Open and Closed Loop Systems, T. W. Sze. *Ph.D. Dissertation*, Northwestern University, Evanston, Ill., Aug. 1954.

7. Analysis and Synthesis of Sampled-Data Control Systems, Eliahu I. Jury. *AIEE Transactions*, vol. 73, pt. I, Sept. 1954, pp. 332–46.

8. Additions to z-Transformation Theory for Sampled-Data Systems, Gladwyn V. Lago. *Ibid.*, pt. II, pp. 403–08.

9. Discrete Compensation of Sampled-Data and Continuous Control Systems, E. I. Jury, W. Schroeder. *Report No. 154*, Electronics Research Laboratory, University of California, Berkeley, Calif., Series 60, 1955.

10. Table of z-Transform and Modified z-Transform of Various Sampled-Data Systems Configurations, E. I. Jury, G. Farmanfarma. *Report No. 136A, Ibid.*

11. Factors in the Design of Digital Controllers for Sampled-Data Feedback Systems, John E. Bertram. *AIEE Transactions*, vol. 75, pt. II, July 1956, pp. 151–59.

12. Digital Compensation of Continuous-Data Feedback Control Systems, K. K. Maitra, P. E. Sarachik. *Ibid.*, May 1956, pp. 107–16.

13. Network Compensation of Error Sampled Feedback Control Systems, J. Ekiansky. *Technical Report T-7/B*, Department of Electrical Engineering, Columbia University, New York, N. Y., April 1, 1955.

Transient Characteristics of D-C Metal-Rolling Motors and Generators

EDWARD P. SMITH
MEMBER AIEE

THE ADVANCE in metal-rolling techniques during the past two decades has been tremendous. The delivery speeds of tandem cold mills have increased from 3,500 to more than 7,000 feet per minute. The reversal time (base speed to base speed) of electric-driven blooming and slabbing mills has been reduced from 2 seconds to less than 1 second with a resultant increase in output. Rod and tube mills have been greatly improved which has resulted in higher quality and a reduction in scrap.

These accomplishments have come about because of a better understanding of the art of metal rolling. This knowledge provides the basis for a more accurate definition of the requirements of the mill prime mover and its associated control equipment. The d-c machine has been broadly applied as the prime mover for metal-rolling mills because of its unique characteristics, of which the most important are as follows: 1. The speed is adjustable over a wide range by varying the armature voltage or the shunt field strength. 2. Motors have high starting torque. 3. D-c machines have high peak load capacity. 4. D-c machines can be easily controlled. These characteristics formulated the basis upon which the d-c machine was applied to mills 50 years ago. During the rapid progress in recent years, the d-c machine has outstripped competitive forms of motive power because these unique characteristics have become even more important to the industry.

The large investment represented in a new rolling mill makes it essential that its performance be accurately determined in advance of its purchase.

With the advent of the complex calculating machines, the rolling mill system performance can be thoroughly analyzed if the characteristics of each component can be accurately predicted. This presents a challenge to the d-c machine designer which must be met.

The steady-state characteristics of the d-c machine are well known. The short-circuit characteristics have been thoroughly studied in recent years, and the results are given in previous AIEE Transactions papers.[1-4] The performance of the d-c machine between these two limits has been analyzed in part by earlier papers.[5,7] Machine constants (used advisedly) have been used in these various analyses. However, if the system studies are to be further advanced, the range of machine constants, the machine characteristics, and the machine performance must be better understood. The purpose of this paper is to provide a ready-reference summary of the transient characteristics of d-c machines and to establish a basis for future developments in this field. This has been effectively done for short-circuit characteristics of d-c machines.[4,8] Corresponding information on transient characteristics will provide additional assistance to estimating machine parameters that are within practical limits of design.

Transient and Subtransient Characteristics

Numerous papers[1-4] have been published in the past several years on the short-circuit characteristics of the d-c machine. In essence these characteristics are of very short time duration and are abnormal in character. However, their importance must not be discounted since proper protection of the machinery from short-circuit damage is obviously essential. The time duration of these transients is a matter of a few milliseconds, whereas the normal and useful duty cycle to which the d-c machine is applied has a time duration of 1/10 second or more. Therefore, the term "transient characteristics" as used in this paper will apply to the load cycle whose period is approximately 1/10 second or more in duration and whose magnitude remains within the peak load capacity of the machine. The short-circuit characteristics of the d-c machine will be identified as "subtransient characteristics," as used for a-c machines.

Paper 56-639, recommended by the AIEE Mining and Metal Industry Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Summer and Pacific General Meeting, San Francisco, Calif., June 25-29, 1956. Manuscript submitted March 29, 1956; made available for printing April 17, 1956.

EDWARD P. SMITH is with the General Electric Company, Schenectady, N. Y.

Direct or Main Axis Field Performance

The total direct axis excitation is important only in so far as it produces direct axis flux which controls the voltage output of a generator or the speed of a motor. This excitation may come from one or more windings of the d-c machine, such as the shunt field, (a single or multiple coil per pole) series field, direct axis effect by half-tune commutation winding, and offset compensating field windings.

In many applications, such as tandem cold mill motors,[8] it is desirable that the flux be determined by the field excitation alone. Most d-c machines have a slight amount of compounding effect from the armature windings, as the result of the basic nature of commutation. On small and low-duty machines this can be corrected by brush position. The larger size motors used in metal-rolling mills require careful adjustment for best commutation, and therefore only slight changes in brush position from the position of best commutation can be tolerated. In the shunt-wound machine the direct axis flux is affected not only by the load current, but also by the rate of change of load current and by changes in machine speed.

Under steady-state conditions, conventional armature excited windings, such as the series field or auxiliary shunt field, can be used to compensate for these effects. The compounding effect of these windings has been found to respond at a different rate than that of the armature reaction. This difference is to be expected of course since the relative locations and configurations of these windings result in a widely different magnetic leakage flux path. The use of offset pole face winding has improved the matching of armature reaction. Here the compounding effect is obtained by a shift in the axis of the pole-face winding, and thus establishes a small component of the pole-face ampere-turns that is effective along the direct axis of the machine. The compounding ampere-turns of this winding are distributed in very nearly the same manner as those of the armature and are also located essentially in the same position in the magnetic circuit. Thus very close matching can be obtained with skillful design for all values of flux, speed, load, and rate of change of load.

It is not practical to design an offset compensating winding as an adjustable device. Instead it must be considered as a means of building a definite amount of compounding into the machine, and for those cases where extremely close steady-state regulation is required, the

offset pole-face winding may have to be supplemented by a conventional adjustable winding, such as an auxiliary shunt field. It has been found that on typical designs, more than 80% (per cent) of the compounding ampere-turns can usually be furnished in the offset pole-face winding with the remainder in the adjustable winding.

This useful design tool has been employed since 1933 to advance the application of d-c machines to more and more difficult jobs. The continued use of this method of compounding since 1933 has given the designer the experience necessary to match more accurately the compounding effect of armature reaction.

The shunt field current of the common shunt d-c machine so compensated is the important winding that controls the direct axis flux. The open-circuit time constant of the shunt field (unsaturated) is the principal constant necessary to predict the transient performance of the machine in this axis. Typical values of this constant are shown in Table I.

Table I. Open-Circuit Time Constant of Shunt Field

Class of Machine	T_{fo} (at 75 C), Seconds
Generator: synchronous motor-generator sets, 40 C open	2.0 to 4.0
Motors: standard, 40 C open, forced ventilated	3.5 to 7.0
Motors: rod mill, 40 C open	3.0 to 6.0
Motors: heavy-duty reversing, 50 C forced ventilated	5.0 to 10.0
Motors: low inertia, 60 C forced ventilated	3.5 to 6.0

The time constants in Table I are based upon the air-gap line of the machine. The design parameters that affect the magnitude of this constant are expressed in the following equation as derived in Appendix I.

$$T_{fo} = 0.0156 K_s^2 B_{g1}^2 K_l \frac{r}{h \Delta T} g_e \frac{P_a}{\left(1 + \frac{w_p}{l}\right)} 10^{-4} \quad \text{seconds}$$

where

K_s = ratio of full load AT/pole to air gap AT/pole no load of shunt field at basic speed

K_l = pole leakage flux linkage coefficient

B_{g1} = air-gap flux density at no load, rated voltage and base speed; lines per square inch

r = shunt field thermal resistance per unit of surface area; degrees centigrade per watt per square inch

h = radial length of shunt coil; inches

ΔT = temperature rise; C

Pa = pole arc; inches

w_p = width of main pole body; inches

l = stacked length; inches

g_e = effective radial length of air gap; inches

The factors that increase the open-circuit time constant for a given thermal rating are

1. High level of saturation in the magnetic circuit.
2. Large magnitude of drooping regulation.
3. High air-gap flux density resulting from large ratio of iron to copper in rotor coil tooth section and from item 1.
4. Ineffective ventilation.
5. Ineffective heat transfer surface of the coil and poor heat conduction within coil; i.e., coil design.
6. Large effective air gap.
7. Large pole arc; i.e., few number of poles and minimum commutation zone.
8. Shortened radial length of shunt coil.
9. Low allowable temperature rise.
10. Large stacked length as compared with pole body width.
11. Large amount of main pole leakage flux linkages.

The design of the machine to meet certain special requirements changes the magnitude of this constant. The increase in temperature rise (ΔT) of the shunt field decreases the time constant proportionally. The maximum permissible temperature rise is limited to some specified value by the economical design and industry standards. The only method of increasing the response of the shunt field circuit without altering some other machine characteristic is to increase the forcing voltage applied to the circuit or to increase the circuit resistance.

If the temperature rise is increased, the forcing voltage may be reduced to maintain the same rate of change of shunt field current. This increases the power consumed in the shunt field, but does not reduce the power rating of the excitation system. Therefore, it is unnecessary to have high-temperature-rise shunt fields for fast-response shunt field circuits. Thus, the most economical machine design should be used, and the excitation system should be designed to drive this circuit properly.

Generators used in metal-rolling mills as power supplies are designed for maximum reliability with minimum initial and maintenance cost. This results in short stacked length, medium- to high-ratio of copper to iron, lower saturation level, effective ventilation, and no speed range, and thus it is to be expected that the time constant should be lower than for motors whose speed and physical configuration are determined by the driven load and its location.

The heavy-duty reversing motors are stacked long to reduce WK^2 and are high-torque devices, which results in a design

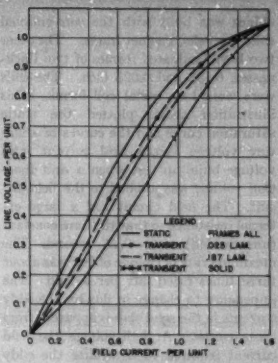


Fig. 1. Transient saturation data for three different magnet frames on duplicate 1,750-kw 514-rpm 600-volt d-c generators. Square-wave field voltage impressed at no load. Field circuit time constant, 0.75 second

in which the ventilation is inherently poorer and the flux level is high. Thus, even though the allowable temperature rise is higher than the 40 degrees allowed for open motors, the other factors increase the time constant.

The shunt field could be conveniently represented by a single circuit constant if its magnetic circuit were free from short-circuited windings. The standard construction of the d-c machine provides a magnetic circuit that is composed of an air gap and an iron portion, partly laminated and partly solid. The solid magnet frame, pole piece punching rivets and keys, the adjusting shims, and the metallic coil form provide media in which eddy currents can be produced by changing flux in the direct axis. The combined effect of these eddy currents oppose the action of the changing shunt field current and a delaying action results in the build-up of the direct axis flux as well as an increase in shunt field current for the same output flux. The major component of the magnetic circuit producing eddy currents is the solid magnet frame. In certain designs the metallic coil form may be an important factor in producing eddy currents, but in these cases the coil form can be eliminated by pole-wound coil or by insulating the form to keep it from forming a short-circuited winding within itself or with the pole piece punching.

The relative effect of magnet frame design is demonstrated in Fig. 1 which plots the test results of three 1,750-kw 600-volt d-c generators, identical except for magnet frame design. One of the ma-

chines was built with the conventional solid-steel-slab magnet frame. The other two have laminated frames of two thicknesses, 0.188 and 0.025 inch. The coils were wound on metal coil forms. This illustration shows plotted the static saturation curve and the curves of armature voltage versus field current from voltage build-up test when a unit function voltage was applied to the field circuit. The field circuit had a large external resistance so that the time constant (unsaturated) was 0.75 second. The maximum rate of change of flux was about three times rated flux per second. The impedance to change in flux by the eddy currents in the solid slab is apparent from these curves. At half-voltage the field current required to neutralize the eddy currents is 75% of that required to produce the flux. Obviously this places a considerable increase in duty upon the excitation system.

As the techniques of rolling mills have advanced, the response of the excitation systems has increased so that the effect of the magnet-frame eddy currents can no longer be neglected or approximated by a simple experience factor. The response of a d-c machine to impact excitation has been published.[10] This should prove to be another tool permitting better representation of the d-c machine under transient operations and should also permit the system designer better to fit the machine to its application.

Quadrature Axis or Interpolar Axis Performance

The armature circuit inductance measurement and method of calculation has been reported thoroughly in recent literature.[9] The machines whose inductances are listed in reference 5 are typical metal-rolling-mill motors and generators. There are four broad classes of machines in so far as armature circuit inductances can be divided:[5]

	Approximate Per-Unit Inductance (C_x)
Motors:	
Without pole-face windings	0.4
With pole-face windings	0.1
Generators:	
Without pole face windings	0.6
With pole face windings	0.2

where

$$L_a = \frac{19.1\ V_t}{P N_1 I_a} C_x \text{ henrys}$$

L_a = armature circuit inductance
V_t = rated voltage
P = number of poles
N_1 = base speed in rpm
I_a = rated armature current in amperes

All of these quantities can be obtained from the machine name-plate data.

The armature circuit inductance is largely determined by the machine rating. The magnitude of this inductance can be varied by a small amount economically without affecting the machine rating.

These factors reduce the inductance as follows:

1. Reduce commutating field coil leakage flux by geometric configuration of the core and coil.
2. For compensated machines obtain 100% compensation of armature reaction by pole-face winding.
3. Reduce slot leakage flux of armature and pole-face bars.
4. Increase the number of circuits in both rotor and stator whenever practical.
5. Use commutating aids such as shading coils, etc.

The armature-circuit time constant is defined as the ratio of armature-circuit inductance to the armature-circuit resistance.

The armature-circuit resistance can be varied over a considerable range without appreciably affecting the inductance. If the machine is designed with normal flux proportions, then as the resistance is increased there is a small increase in inductance. The armature resistance is one of the major factors in determining the speed regulation of a motor.

The per-unit magnitude of this resistance depends principally upon these factors: the base speed, the speed range, the maximum overload, the voltage, the temperature rise, the steady-state regulation, the effectiveness of ventilation, the type of compensation, and the length-to-diameter ratio. Therefore, this parameter varies over a considerable range depending upon the application.

Some typical values are given in Table II. These are based upon 75 C winding temperature and include a 3-volt brush drop.

Table II

	Armature Circuit	
	Resistance per Unit (r_a)	*Time Constant, Seconds (t_1)
Generators: 40 C, 200% maximum load, synchronous motor generator sets	0.02 –0.03	0.02 –0.025
Motors: standard, 40 C open	0.025–0.06	0.02 –0.04
Motors: rod mill, 40 C open	0.02 –0.05	0.03 –0.07
Motors: low-inertia, 60 C forced ventilated	0.05 –0.09	0.025–0.04

*All machines are compensated, these values are increased approximately three to four times for uncompensated machines.

Many of the factors that affect the machine resistance also affect the magnitude of the inductance. The range of armature-circuit time constant for one particular class of machines is smaller than the range of resistance.

From a comparison of the magnitudes of the armature time constant with the shunt-field time constant, it is apparent that the shunt-field time constant is approximately 100 times larger. In many applications the armature-circuit delay may be neglected without affecting the accuracy of the results. However, on applications utilizing high speed of response regulating systems as used on rod and tube mills,[6,7] this constant is most important in determining the transient performance.

Mechanical or Speed Axis Performance

The third axis of performance of the d-c motor evaluates the ability of a motor to develop torque from the stored energy in the rotational mass with loss in speed and the dissipation factors connected therewith. This time constant is $t_s = 2\ H\ r_a$ (see Appendix II) where H is the per-unit inertia consant and r_a is the per-unit resistance of the armature circuit.

The armature resistance of the motor varies from 2 to 9% (see Table II). The inertia constant of the motor alone can vary over a wide range depending upon the design of the motor. The values given in Table III are typical of modern-design motors of 2-to-1 speed range and 200% maximum load. The values given in Figs. 2 and 3 have been developed from typical modern designs. These data have an accuracy of ±10% from actual design for all National Electrical Manufacturers Association class N and S ratings, and should prove valuable in making initial evaluations of various system designs.

Table III

Type of Motor	H at Base Speed
40 C open	0.30–1.00
40 C open (rod mill)	0.60–1.20
40 C forced ventilated (standard)	0.35–0.60
40 C forced ventilated, low inertia (see Fig. 2)	0.20–0.40
50 C forced ventilated, low inertia (see Fig. 3)	0.15–0.30
60 C forced ventilated, low inertia	0.10–0.20

Other requirements of the motor will affect the design to increase these ranges, such as a wider speed range, heavier overloads, space limitations, and mechanical combinations of double and triple

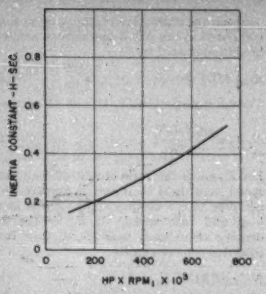


Fig. 2. Inertia constant H for forced-ventilated low-inertia motors rated 40 C, 200% maximum load, 600–750 volts, 2-to-1 speed range

armatures in tandem. In metal-rolling mills, except for tandem cold reduction mills, the load inertia is usually considerably less than the motor inertia. However, on all mills the inertia of the drive motor is of major importance.

To illustrate the latitude the machine designer has in controlling the inertia of a d-c motor, the following basic design equations will be used. The torque equation is well known and it applies to all rotating electric machines.

$$\text{Torque} = \frac{\text{hp}}{\text{rpm}} = K_1 D^2 L \qquad (1)$$

where

D = diameter at the air gap
L = effective length of the core
K_1 = a constant
rpm = base speed value for variable-speed motors

K_1 is a constant only for a limited range of machine sizes of the same type. The machines of this size range must have the same temperature rise, ventilation system, peak overload, and mechanical construction.

This equation applies to synchronous and induction motors as well as d-c machines except that the value of K_1 may be different. The ratio of D to L in the synchronous and induction machine is determined by mechanical and economic limitations. The d-c machine has two other design limitations that are a direct result of commutation and commutator design. The first of these is the limitation of the maximum voltage between adjacent commutator segments. Since it is necessary to bring the coil leads out to the commutator segments, the maximum voltage that can be generated per coil is limited by the voltage that can be sustained across the creepage surface of the side mica. Therefore, the average of peak volts per bar on the commutator is limited to a rather low value in order to minimize the possibility of arcing between segments and commutator flashovers. The other design limitation is the reactance voltage of the coil undergoing commutation. Experience has shown that on machines of similar design, the reactance voltage is the criterion for successful commutation.

Both of these limitations result in the second basic design equation that is unique with d-c machines.

$$\text{Minimum gap diameter} = K_2 \text{ hp} \qquad (2)$$

where K_2 is a constant similar to K_1 in that it applies only to machines rated on the same temperature rise, ventilation system, speed range, peak load, type of armature winding, and mechanical construction. If the limiting value of volts per bar is increased, the gap diameter can be decreased. A higher speed range, lower temperature rise, and less effective ventilation all increase the gap diameter for the same horsepower output.

The volume of material to obtain the torque is given by equation 1. The inertia varies as the fourth power of the diameter of any homogeneous cylindrical solid. The rotor of a motor follows this basic law. Therefore, the inertia can be reduced by reducing the diameter until equation 2 limit takes effect. However, the mechanical limitations must be considered. If equation 2 is applied to determine minimum gap diameter of motors with low horsepower and speed but with high torque, the stacked length L (from equation 1) may become excessive. In this case, it may be impossible to ventilate the rotor, and so a rotor diameter larger than minimum must be used.

The high-speed motor has its diameter limited by the centrifugal force on the windings. This force is approximated by the peripheral velocity which is the product of rpm $\times D$. Then from equation 2, it follows that the product of hp $\times$ rpm (top speed) is limited to a maximum value. A practical limit of this product is 2.5×10^6. This value can be increased by increasing the peripheral velocity above the economic and practical limit of 10,000 feet per minute, which normally requires a more costly design and usually results in a higher maintenance machine.

Commutation

A discussion of transient performance of d-c machines would be incomplete if the subject of commutation was not included. As the generators are high-

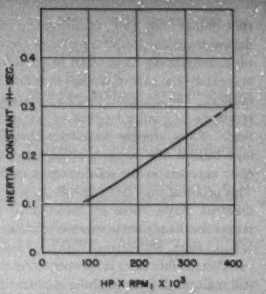


Fig. 3. Inertia constant H for reversing hot mill motors rated 50 C, 275% maximum load 600–900 volts, 2-to-1 speed range

speed machines, commutation is the factor that limits the operating speed. The low-speed motor may have commutating factors which are limiting if the design is based on minimum armature diameter as required for minimum shaft spacing or low armature inertia.

The principal factors influencing successful commutation for steady-state operation are the reactance voltage, correct compensation both in phase and magnitude, adequate neutral, linear flux current characteristic of commutating pole, proper brush material, adequate armature winding equalization, riding qualities of the brush on the commutator, and atmospheric conditions. All of these factors must be properly balanced for successful commutation on steady-state loads.

The same factors are important for successful commutation under transient loads. However, under transient operation, the armature current, the speed, and the direct axis flux are changed independently or simultaneously. The coil undergoing commutation links the direct axis flux so that a voltage is induced in this coil by transformer action of the changing flux. This voltage adds to the reactance voltage of this coil. The commutating pole flux is adjusted to compensate for only the reactance voltage, therefore this transformer action voltage reduces the margin of commutation. In most variable-voltage applications in recent years this transformer voltage has not been large enough to affect commutation. But with the advancement of rolling techniques increased rate of change of armature voltage will be required, and this voltage may become a limiting factor in successful commutation.

The commutating field flux must maintain its linear relation with armature cur-

rent under all load conditions. Thus during transient loads, the rate of change of commutating field flux must follow accurately the rate of change of armature current. The commutating field magnetic structure is similar to that of the main field. Thus, during transients this flux lags behind the current because of the eddy currents in the solid magnet frame. This problem may be overcome in several different ways. One of these is to increase the steady-state margin of commutation, thus permitting some transient variation in flux from its proper value and still maintaining successful commutation. This may be accomplished by increased armature diameter, using high-reluctance commutating field design. and proper proportioning of the magnet frame to minimize the eddy-current effect. Another solution is to laminate the magnet frame. The method used by Brockman and Linkous[10] can be extended to the commutating field circuit. This provides the designer the tool to determine in advance the most economical design that will provide successful commutation.

Conclusions

1. The d-c machine has three principal parameters that describe its transient behavior within a given system. These parameters can be considered constants under defined conditions.

2. The magnitude of these constants has a wide range of values for different ratings within a given class of machines.

3. Because of the eddy currents in the solid magnet frame during transients, a secondary set of parameters is necessary to define the transient behavior of the direct and quadrature axes.

Appendix I. Factors Affecting Shunt Field Time Constant

The unsaturated inductance of shunt field (assuming all poles are connected in series) is defined by:

$$L = \frac{\phi N_f P}{I_{f0}} K_l \times 10^{-8} \qquad (3)$$

where

$\frac{\phi}{I_{f0}}$ = slope of air-gap line of saturation curve, flux lines per pole per ampere
P = number of poles
N_f = number of turns per pole
K_l = pole leakage flux linkage coefficient
L_{f0} = inductance in henrys

The average flux density over the surface of the armature is expressed as

$$B_{av} = \frac{P\phi}{\pi D_a l} \qquad (4)$$

where

B_{av} = average flux density, lines per square inch
D_a = rotor diameter, inches
l = length of over laminations, inches

If α is the ratio of the pole arc to the pole pitch and B_g is the average flux density under the pole arc, then

$$B_{av} = \alpha B_g \qquad (5)$$

The ampere-turns to produce this average flux density across an effective air gap of g_e is

$$N_f I_{f0} = 0.313 B_g g_e \qquad (6)$$

Combining these for relationships results in

$$L_{f0} = \frac{\alpha D_a l N_f^2 K_l}{g_e} \times 10^{-7} \qquad (7)$$

By definition, the shunt-field open-circuit time constant (unsaturated) is

$$T_{f0} = \frac{L_{f0}}{R_f} \qquad (8)$$

where

R_f = resistance of the shunt field in ohms

The resistance of the shunt field is determined by

$$R_f = \frac{\rho P N_f MLT}{A_c} \qquad (9)$$

where

ρ = resistivityof winding conductor material, ohm-inches
MLT = length of turn, inches
A_c = cross-section area of one turn, square inches

By combining equations 7, 8, and 9, the time constant is expressed in turns of the physical parameters of the machine. Thus

$$T_{f0} = \frac{K_l \alpha D_a l N_f A_c}{\rho P MLT g_e} 10^{-7} \qquad (10)$$

The temperature rise of the shunt field depends upon the heat-transfer coefficient, the surface area, the loss in the winding. This may be expressed as

$$\Delta T = \frac{r}{PA_s} I_f^2 R_f \qquad (11)$$

where

ΔT = temperature rise, C
r = total thermal resistance per unit area, C per watt per square inch
A_s = heat-dissipating surface area of coil, square inch
I_f = rated field current, amperes

On similar types of construction the outer surface of the coil may effectively be used for the heat-dissipating surface of the coil if the equivalent thermal resistance per unit area is used. Therefore

$$A_s = \left(MLT + \frac{\pi w_c}{2}\right) h \qquad (12)$$

where

w_c = thickness of the coil, inches
h = radial length of the coil, inches

Since the thickness of the coil is small compared to the MLT, the surface area can be approximated by

$$A_s \doteq MLT\, h \qquad (13)$$

Combining equations 9, 11, and 13 results in

$$N_f A_c \doteq \frac{\rho r (N_f I_f)^2}{\Delta T h} \qquad (14)$$

Let K_s equal the ratio of the full-load base-speed shunt-field excitation in ampere-turns per pole to the air-gap ampere-turns at no load and base speed. If B_{g1} is the flux density at no load, base speed, then from equation 6 there results

$$N_f I_f = 0.313 K_s B_{g1} g_e \qquad (15)$$

The mean length of turn can be approximated by twice the sum of the stacked length and the width of the pole body, thus

$$MLT \doteq 2l\left(1 + \frac{w_p}{l}\right) \qquad (16)$$

where

w_p = width of pole body, inches

By combining equations 10, 14, 15, and 16, the open-circuit time constant may be expressed as

$$T_{f0} \doteq 0.49 K_s^2 K_l B_{g1}^2 \frac{r}{h\Delta T} g_e \frac{\left(\frac{\alpha D_a}{P}\right)}{\left(1 + \frac{w_p}{l}\right)} \times 10^{-8} \qquad (17)$$

By definition the pole arc Pa is $\pi a D_a / P$. So, using this identity,

$$T_{f0} \doteq 0.156 K_s^2 K_l B_{g1}^2 \frac{r}{h\Delta T} g_e \frac{P_a}{\left(1 + \frac{w_p}{l}\right)} \times 10^{-8}$$

seconds (18)

Appendix II. Speed Axis Time Constant

This constant is derived in the analysis of the speed transient of d-c mill motors.[4,7] The defining equation for this constant is

$$t_s = 0.46 \times 10^{-8} \frac{WK^2 R}{K_v^2} \qquad (19)$$

where

t_s = speed axis time constant, seconds
WK^2 = inertia of rotor, pound-feet^2
R = effective armature circuit resistance, ohms
$K_v = V_b / N_b$ = armature volts per rpm
N_b = basic speed in rpm

The per-unit armature circuit resistance is defined by

$$r_a = \frac{I_b}{V_b} R \qquad (20)$$

where

I_b = rated armature current, amperes
V_b = rated armature voltage, volts

The inertia constant of a machine H_b is defined as the ratio of the stored energy in

the rotor at basic speed to the machine rating. This is expressed as

$$H_0 = \frac{0.231 WK^2 N_1^2}{KW_0} \times 10^{-6} \text{ seconds} \qquad (21)$$

or

$$= \frac{0.231 WK^2 N^2}{V_0 I_0} \times 10^{-3}$$

Thus, by combining equations 19, 20, 21, there results

$$i_0 = 2H\sigma'_0 \qquad (22)$$

References

1. Current and Torque of D-C Machines on Short Circuit, T. M. Linville. *AIEE Transactions*, vol. 65, 1946, pp. 956–65.

2. Solid Short Circuit of D-C Motors and Generators, T. M. Linville, H. C. Ward, Jr. *Ibid.*, vol. 68, pt. I, 1949, pp. 119–24.

3. Transient Characteristics of D-C Motors and Generators, A. T. McClinton, E. L. Brancato, Robert Paneff. *Ibid.*, pt. II, pp. 1100–06.

4. Maximum Short-Circuit Current of D-C Motors and Generators, AIEE Committee Report. *Ibid.*, vol. 69, pt. I, 1950, pp. 146–49.

5. Rate of Rise of Short Circuit Current of D-C Motors and Generators, AIEE Committee Report, A. G. Darling, T. M. Linville. *Ibid.*, vol. 71, pt. III, Jan. 1952, pp. 314–25.

6. Speed Transients of D-C Rolling Mill Motors, L. A. Umansky, T. M. Linville. *Ibid.*, vol. 54, Apr. 1935, pp. 387–94.

7. An Extension of Impact Speed-Drop Analysis, F. E. Crever. *Ibid.*, vol. 66, 1947, pp. 191–96.

8. Developments in the Design of D-C Motors and Generators for Rolling Mills, B. E. Caldwell, Jr. *Iron and Steel Engineer*, Pittsburgh, Pa., Dec. 1950.

9. Measurement and Calculation of D-C Machine Armature Circuit Inductance, H. D. Snively, Paul B. Robinson. *AIEE Transactions*, vol. 69, pt. II, 1950, pp. 1228–37.

10. D-C Machines: Response to Impact Excitation, J. J. Brockman, C. E. Linkous. *AIEE Transactions*, vol. 74, pt. III, June 1955, pp. 500–05.

Discussion

E. H. Myers (Westinghouse Electric Corporation, East Pittsburgh, Pa.): Mr. Smith has effectively summarized many of the factors affecting the "transient characteristics of d-c machines. It is hoped that his suggested definition of the terms "transient characteristics" and "subtransient characteristics" will be adopted generally for d-c machines.

The designer of large d-c machines for metal working is fortunate in having customers whose insistence on quality permits him to design for the best "tailor-made" machine possible for a particular application. Cost is secondary. Machine performance and mill output are primary goals.

On this equipment, the general rule is followed that, whenever possible, the cause of possible misoperation is eliminated rather than an attempt to overcome the results by working on other factors. A good example of this is the main generator for slabbing and blooming mills. It has now become general practice for these high-speed high-reactance-voltage generators to be built with laminated frames instead of solid rolled steel slabs. This prevents eddy currents being set up in the frame structure which would delay the commutating pole useful flux. Any delay in this flux behind the reactance voltage of commutation which is in phase with armature current changes results in severe sparking. Every means is used to reduce the steady-state reactance voltage to a minimum but in addition the use of thin electrical sheet-steel laminations approximately 0.018 inch thick reduces the frame time constant by a factor of more than 100 to 1.

Even with the freedom "to do his best," however, the designer finds many of his design tools that solve some problems raise other ones. Design still remains an art of compromise. Long-core small-diameter motors have low inertias and reduced peak loads during acceleration and deceleration but unfortunately also have high commutation reactance voltages, high voltage between commutator bars, and are more difficult to cool. Large-diameter machines have low-reactance voltage of commutation and low impact speed drops but may have high commutator peripheral speeds and require more power to accelerate and decelerate. A large number of carefully spaced compensating bars result in excellent compensation, better regulation curve-shapes, and lower voltage between commutator bars, but also tend to be more subject to distortion during overloads due to cobbles and this design tends to use up valuable pole area that could be used for carrying useful flux. A few heavy compensating bars are far more sturdy. The offset compensating winding is an economical source of series-compounding action but is fixed in position and magnitude. The designer must decide if the possible disadvantages of improper compensation at the pole tips, the restriction to one direction of rotation, and lack of complete adjustment outweigh the advantages.

The very thorough study of the factors affecting the shunt-field time constant outlines many of the design considerations investigated by engineers in first setting up design standards for their company. Assuming that the best cooling methods have been long established, the problem of designing a shunt field for the lowest time constant is reduced to that of obtaining an acceptable field with the absolute minimum pounds of copper in it and in eliminating any unnecessary damping. Of far greater importance in deciding the speed of flux changes is the external excitation circuit.

Mr. Smith's tables should be of most interest to those not directly concerned with actual machine design. The approximate values given may be used as guides to those making preliminary studies of possible drive arrangements. However, the demands of modern drives are so exacting that such general studies must be carefully confirmed using actual machine data for the particular equipment involved. Most machine factors are closely interrelated and final results often cannot be predicted with accuracy without actual machine designs.

E. P. Smith: Mr. Myers presents the problems that confront the designer in "tailor-made" machine design. Past experience of the designer will affect his "art of compromise." Mr. Myers points out that costs are secondary in this type of equipment. I feel that a more satisfactory criterion is successful machine performance at minimum cost. This is the best design for both the builder and the user.

Mr. Myers emphasizes the principal purpose of this paper. That is, to provide guiding data for those making preliminary studies of the possible drive arrangements using d-c machines. The final design of a metal-rolling mill system must be based upon actual machine data. I am in full agreement with Mr. Myers' comment.

The excitation system is most important in the control of the direct axis flux; however, the response of this system is dependent, not only upon the shunt-field time constant, but on the power required by the shunt field. In the field of electronics, a power amplifier has a very significant criterion measured by a term called the figure of merit. This term is defined as the ratio of the amplification factor divided by the time constant. A d-c generator or exciter may be evaluated by using this same parameter. Using Esson's output equation of a d-c machine and the various relations in Appendix I of the paper, it can be shown that the figure of merit is related to five principal parameters of the d-c generator. Esson's equation is $KW = KB_g \sigma a D_a^2 l N_1$

where

KW = output of d-c machine
K = constant
B_g = average flux density in gap
σ = armature loading, ampere conductors per inch of periphery
α = ratio of pole arc to pole pitch
D_a = armature diameter
l = length over laminations
N_1 = rated base speed

The amplification factor of a d-c machine is defined as the output kilowatt divided by the kilowatt loss in the shunt field winding. This results in the figure of merit being equal to:

$$\text{Figure of merit} = \frac{K\sigma D_a N_1}{K_1^2 K_2 g}$$

The best figure of merit is obtained when a generator or exciter has the following:

1. A high peripheral velocity ($D_a N_1$).
2. Maximum armature loading or a large number of armature conductors.
3. Low saturation.
4. Minimum shunt field leakage flux.
5. Small air gap.

Most of these factors result in an increased cost of the machine if other performance characteristics, such as maximum load commutating ability, high efficiency, and conservative temperature rises such as 40 C, are maintained. The proper balance between the figure of merit and the excitation circuit external to the machine must be obtained for the best over-all system performance.

Centralized Control of Power Systems in the Steel Industry

W. A. DERR
MEMBER AIEE

W. L. METZ
ASSOCIATE MEMBER AIEE

OPERATING economies and efficiencies are realized from the concentration of power system control and indication in a single central location. The average steel works power system is ideal for the application of centralized control equipment.

The electric power for a steel works may be provided by steel company generators, it may be purchased from power companies, or it may be a combination of owned and purchased generation. The consumption of relatively large amounts of power at various locations in the works area necessitates the use of a number of substations to provide for efficient and adequate routing of power. The transmission and distribution substations involved in supplying power to the proper load centers in a steel works are similar to the transmission and distribution substations of power companies. However, there is a decided concentration of such substations in a relatively small area compared with the usual configuration of such substations of a power company. This means that a central control location can be chosen for all substations of a steel works which is a relatively short distance from all substations. When there is local steel company generation, it is usually preferable to have the central control location at or near the power house. Fig. 1 shows the concentration of substations at the Geneva (Utah) Works, Columbia-Geneva Steel Division of the United States Steel Corporation.[1] Centralized control of all of these substations is provided from a control center located adjacent to the powerhouse.

Remote Control Equipment

Power-line carrier or microwave channels are often employed for high-speed protective relaying and for remote control of widely separated electrical utility company substations. However, the relatively short distances between steel works substations dictate the use of metallic circuits for high-speed protective relaying and remote control. Wires in underground telephone-type cables are usually employed for pilot-wire relaying for high-speed clearing of faults on the transmission lines between substations and for centralized control.

When metallic conductors are provided for remote control, there can be a choice between direct-wire control and supervisory control. Systems which function by direct energization of control and indication circuits over the metallic conductors between the remote substation and the control center are known as direct-wire control systems. With any direct-wire control system, the number of wires required between the control center and a remote substation is proportional to the number of control and indication functions. Supervisory control systems function by transmitting coded signals between the control center and the remote substation over a relatively few number of conductors. In contrast to all direct-wire control systems, modern supervisory control can provide for an unlimited number of control and indication functions over a single pair of wires between the control center and the remote substation.

In addition to requiring a minimum number of conductors between the control center and the remote substations, supervisory control provides the following important advantages over any direct-wire control system:

1. The control wires are continuously supervised for open or short circuits, thus making the control center operator immediately aware of loss of control. Line conductor supervision is not practical with any direct-wire control system.

2. Supervisory control equipment can provide for future additions to the system without the necessity of providing additional line wires, and with a minimum of changes.

3. Protection of terminal equipment from high voltages on line wires is simplified with supervisory control because of minimum number of line wires required.

4. A separate telephone system is not required, since voice communication can be provided as part of the supervisory control system.

5. Selective telemetering can be provided by supervisory control without the use of additional conductors.

Most existing centralized control of steel works power systems has been provided by supervisory control.

Functions Performed

To provide adequate centralized control of a power system, it must be possible to perform all essential control functions at the individual substations from the control center. Likewise, position-indicating lamps must continuously supervise the position of devices involved in the functioning of the substations.

Continuous or selective telemetering of essential quantities must be provided at the control center. Telephone service between the control center and each remote station for emergency communication is essential.

The choice of controls, lamp indications and quantitative information to be provided is the most important decision made in the application of a centralized control system.

Circuit Breakers and Motor-Operated Switches

The most important items involved in the centralized control of any power system are the circuit breakers. It is customary to provide for the control of all circuit breakers in all substations from the control center with the position of all of the circuit breakers being continuously indicated by red and green lamp indications at the control center. Motor-operated switches which are important to the system operation are also usually controlled and supervised from the control center.

Synchronizing of some of the substation breakers of a steel works power system is often desirable for best system operation. Fig. 2 shows the primary power system of the Fairless Works at Morrisville, Pa.[2] The four substations shown are controlled from the powerhouse. It is apparent from this single-line diagram that synchronizing of all of the 69-kv breakers at the four substations is essential for efficient system operation. Where such synchronizing of remote circuit breakers is essential or desirable, a synchronism indication must be provided at the control center.

It is also necessary that the operator at the control center have direct or indirect control of the frequency of one of the two

Paper 56-731, recommended by the AIEE Mining and Metal Industry and Industrial Power Systems Committees and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Summer and Pacific General Meeting, San Francisco, Calif., June 25-29, 1956. Manuscript submitted March 28, 1956; made available for printing May 1, 1956.

W. A. Derr and W. L. Metz are with the Westinghouse Electric Corporation, East Pittsburgh, Pa.

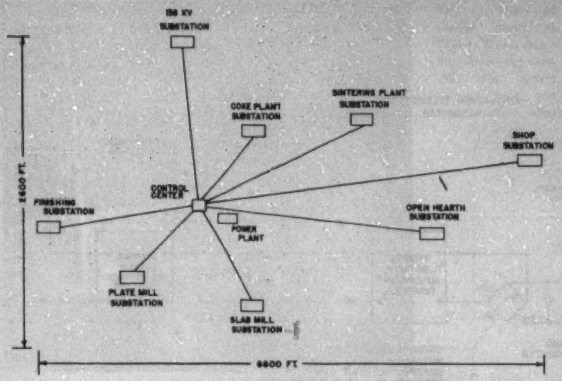


Fig. 1. Approximate geographic layout of control center and remotely controlled substations at Geneva Works of United States Steel Corporation

sources to be tied together. In the case of the system shown in Fig. 2, the operator has direct control of the generators in the powerhouse and can make the necessary adjustments to bring the two systems into synchronism.

It is customary to interlock the control of circuit breakers which must be synchronized so that a "synchronizing" push button must be operated at the control center before a control operation can be performed. The synchronizing indication provided can be obtained in two ways. Since the distances involved are relatively short, and since cables are usually used to provide the necessary line wires for supervisory control and pilot-wire relaying, additional cable wires are often provided to connect the proper potentials directly from the substation to a standard synchroscope at the control center. Potentials are only connected to the synchronizing line wires when a circuit breaker is selected by the supervisory control prior to operation.

The other method of obtaining a synchronism indication is to employ a synchronism telemetering transmitter which has a d-c output and which can be operated over the supervisory control line wires, therefore eliminating the need for any additional conductors. Fig. 3 shows the circuitry of the synchronism telemetering transmitter together with the zero-center milliammeter which is used to provide the indication of synchronism.

As shown in Fig. 3, the single-phase potentials to be synchronized are connected to synchronous motors *M1* and *M2* and transformers *T1* and *T2* of the synchronism telemetering transmitter. The small synchronous motors *M1* and *M2* are coupled together through differential gearing to operate a double-pole double-throw switch. If the frequency of the voltage connected to motor *M1* is higher than the frequency of the voltage connected to motor *M2*, contacts *M1* are closed and contacts *M2* are open. Likewise, if the frequency of the voltage connected to motor *M2* is the higher, contacts *M2* are closed and contacts *M1* are open. Motors *M1* and *M2*, with their associated differential gearing function, detect a frequency difference between the voltages to be synchronized and determine the direction of current flow which is applied to the supervisory control line wires for synchronism indication.

The secondaries of transformers *T1* and *T2* of the synchronism telemetering transmitter are connected in opposition and the vector difference of the two voltages is rectified. This rectified voltage is connected to the line wires by means of the supervisory control equipment and results in the flow of a small direct current through the zero-center d-c milliammeter provided for synchronism indication. The polarity of this current is governed by contacts *M1* and *M2* and the pointer of the synchronism indicator swings between zero and full scale at a rate proportional to the frequency difference. The side of zero to which the pointer deflects indicates which of the two voltages has the higher frequency.

Although the substations are normally controlled from the control center, it is customary to provide motor-operated switch and circuit breaker control switches and indicating lamps in the substations for local control. A manually operated control-transfer switch with "supervisory" and "local" positions is usually provided for each circuit breaker and motor-operated switch. When such transfer means are employed, a circuit breaker or motor-operated switch may be operated

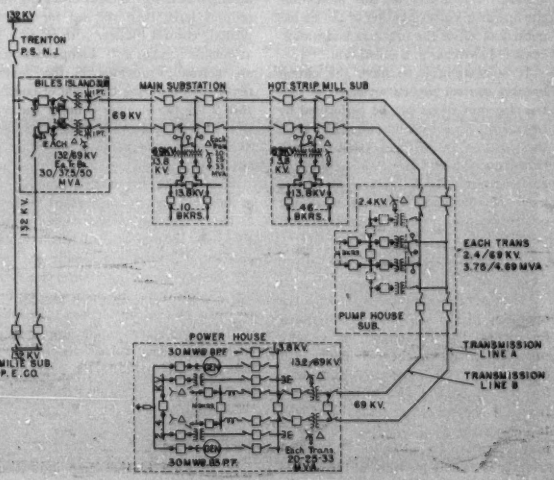


Fig. 2. Primary power system of Fairless Works of United States Steel Corporation

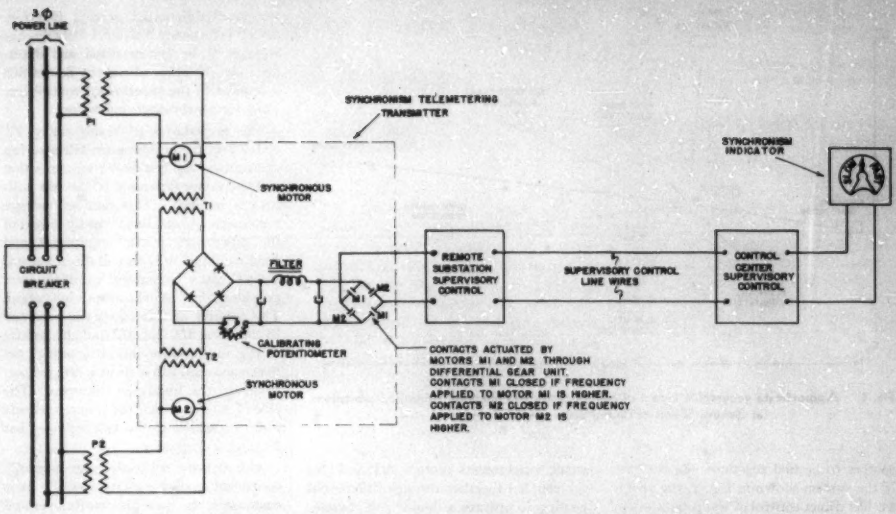


Fig. 3. Diagram of synchronism telemetering transmitter including external connections

from the control center only when its associated control-transfer switch in the remote substation is in the supervisory position. A common indication in the control center is usually provided for all of the control-transfer switches in a substation. By means of this indication, the operator knows whether or not he has control of all of the breakers and motor-operated switches in a substation.[3]

It is customary to have the circuit breaker control circuits arranged so that tripping operations can be performed by the local control switches regardless of the position of the associated control-transfer switches. However, closing operations can only be performed with a local control switch when the associated control-transfer switch is in the local position.

Tap Changers

When purchased power is used for the operation of the steel works, and transformer at the tie station is usually provided with tap-changing-under-load equipment. It is customary to arrange for remote control of the tap changer from the control center even though the tap changer may be equipped with automatic regulating equipment. A control transfer is usually provided at the remote station so that the tap changer can be controlled either locally or remotely, depending upon the position of the transfer switch. If the tap changer is equipped with automatic regulating equipment, a means is provided for transferring between automatic and manual control, both locally and remotely. Preferred arrangements for efficient operating procedures have been devised for accomplishing control transfer with or without automatic regulation.[3] Lamp indications are normally provided in the control center for indicating the point of control, and, if the tap changer is under the control of the control center, then lamps indicate whether the control is manual or automatic.

To control the tap changer remotely one or more of several quantities are required to be telemetered to the control center. Usually it is desirable to telemeter tap position. However, in some cases it is satisfactory merely to supervise with lamp indications the extreme limiting positions of the tap changer. In order to determine the required raise or lower control functions, volts, vars, or power factor, depending upon the quantity to be maintained, should be telemetered to the control center.

Alarm Indications

In addition to the lamp indications showing the position of devices controlled or supervised, it is customary to provide lamp indications of various alarm conditions for most efficient operation of the power system. All alarms need not be individually supervised. Trouble-detecting devices are often grouped together at the remote substation to provide a common alarm indication if any one of the devices in the group detects trouble. A common arrangement consists of including an annunciator in each remote substation with provisions for supervising the annunciator at the control center. All alarms within each substation are then indicated on the local annunciator and reported to the control center as one common alarm for each remote substation. Usual practice provides individual indications for such alarms as substation door, transformer differential lockout, transformer high temperature, or low oil level. It is sometimes deemed desirable also to supervise individually some of the following conditions:

1. Low air pressure for pneumatically operated circuit breakers.
2. Tap changer off-tap indication.
3. Low station battery voltage.
4. Faulty pilot wires for pilot-wire relaying.
5. Bus differential relay operation.

If alarms originating in low-voltage areas not considered a part of the centralized control system are of primary

Fig. 4. View of control center at Fairless Works showing supervisory control desk, system diagram board, and power station control board

Fig. 5. Supervisory control escutcheon

concern to the centralized control operator, then it is desirable to have these alarm conditions indicated in the control center. These indications can be obtained either by direct wire from the point of origin to the control center or by direct wire from the point of origin to the nearest remote substation equipped with supervisory control from where they can be reported to the control center.

Telemetering

Some quantitative indications are desirable and usually required for the efficient operation of remotely controlled substations.

Most telemetered indications are usually provided on a selective or on-call basis whereby the system operator selects the telemetered indications periodically or as required by means of the centralized control system equipment. Since the distances involved in a steel works are short, the necessary quantitative information is often obtained by connecting the associated potential and current transformers to the cable circuits through auxiliary potential transformers and step-down current transformers. Voltage and synchronizing indications are the quantities most frequently obtained in this manner. It is usually more economical to obtain other quantitative information by the use of standard telemetering equipment.

For each quantity for which it is deemed necessary to have a continuous indication at the control center, a separate telemetering circuit is required. Usually there are relatively few quantities which must be telemetered continuously since a fast-acting centralized control system can readily provide the operator with any desired indication. Exceptions concern those quantities for which a continuous recording is required such as the interchange of power at a tie station with an electrical utility company.

There are usually some quantities which it is desirable to have simultaneously telemetered with control operations in order to afford the system operator with the most efficient operating procedure. Those quantities most commonly combined with control in steel works applications are synchronizing indication combined with circuit breaker control in the manner described previously, and tap changer position or an associated quantity combined with the raise and lower control of the tap changer. Other quantities can be combined with control if economic advantages or simplified operating procedures can be realized. As an example, the telemetering of line current is often combined with circuit breaker control. Some or all of the following quantities may be desirable information to have telemetered to the control center from remotely controlled substations.

1. A-c volts.
2. Watts.
3. Vars.
4. Power factor.
5. Tap changer position.
6. A-c amperes.
7. Battery voltage.

Quantities which can be momentarily interrupted are frequently telemetered over the supervisory control channel with the supervisory control equipment at rest. The remote station battery voltage is commonly telemetered in this manner, being

Fig. 6. View of control center at Geneva Works

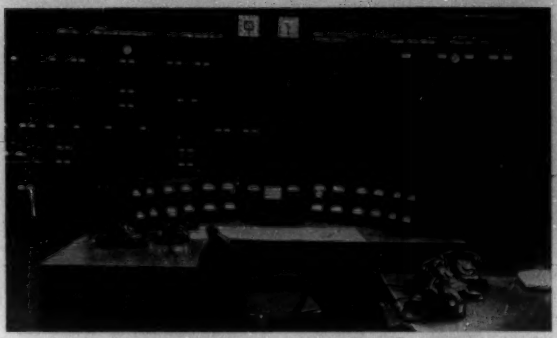


Fig. 7. View of control center having supervisory control escutcheons located in mimic bus layout. Telephone communication to all controlled substations is provided by supervisory control

interrupted only while the supervisory control equipment is in operation.

Communication

Communication between the control center and each remote station included in the centralized control system is usually essential. During emergency conditions or routine maintenance and testing when a man is dispatched to a remote substation, the system operator benefits by having voice communication with the remote station. The communication system can use separate circuits and be entirely separate from the centralized control system. However, when supervisory control provides the centralized control system, it is customary and most economical to include communication as a part of the supervisory control system. The ringing function is then performed by the supervisory control equipment and the supervisory control channel is used as the voice transmission medium. By means of telephone selector keys the system operator can arrange to talk to any one substation or to all remote substations simultaneously, and can also arrange for any combination of remote stations to communicate with each other. If desirable, telephone extensions can be provided at each remote station; for example, a telephone can be provided in the substation control building and a telephone extension can be provided outside in the switchyard.

Physical Arrangements

Control Center

There are many possible physical arrangements of control center equipment. Most of the large installations in the United States include an electrically operated map or mimic bus arrangement which depicts each substation and device within each substation in its electrical position. In so far as practical, the devices are also shown in their correct physical location. The electrically operated map continuously indicates the position of each device under the control of the system operator. All indicating instruments are usually mounted on the electrically operated map, or diagram board as it is sometimes called. The indicating instruments can be located in the mimic bus to indicate the exact electric circuit to which the instrument refers

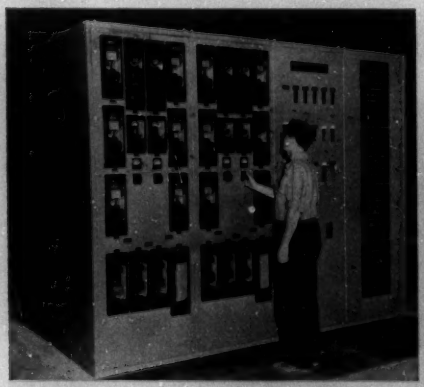


Fig. 8 (above). Duplex control board in remotely controlled steel works substation. Supervisory control relays are on right-hand panel

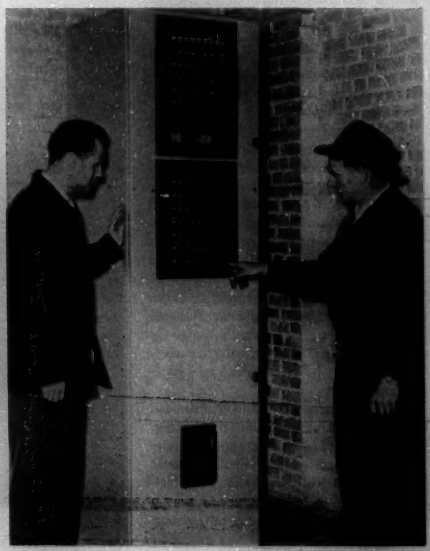


Fig. 9 (right). Supervisory control unit in remotely controlled steel works substation

or they can be located on the upper portion of the diagram board. The instruments are more commonly located above the actual mimic bus layout to conserve space, and to reduce the number of instruments required since where selective telemetered indications are used, one instrument is usually used in connection with several remote station circuits.

The diagram board frequently consists of the front panels of a tunnel-type duplex switchboard. A typical installation of this type of construction is used with the electric system described by Fig. 2, and is shown in the left-hand background of Fig. 4. The right-hand switchboard in this illustration is the control board for the steel company's own power station. The diagram board and the power station control board are of the same construction, both being tunnel-type duplex switchboards. The diagram board includes only the indicating instruments and indicating lamps. All remote control operations are performed from the control desk in the foreground.

The control desk in the foreground of Fig. 4 is a knee-hole furniture-type steel desk on the top of which are mounted an escutcheon for each device being controlled or supervised and the necessary master control push buttons and indicating lamps. The escutcheons are of the type shown in Fig. 5. Each escutcheon of this type has red and green indicating lamps, which are duplicates of those on the diagram board, for indicating the position of the device in the remote station associated with the escutcheon. Each escutcheon also has a push button for selecting the device associated with the escutcheon and a white lamp for verifying that the desired selection has been made prior to performing a control operation. To indicate operations initiated at the remote stations, either the red or green lamp will flash in accordance with the new position of the device being supervised until recognized by the system operator.

A modified arrangement of the layout of Fig. 4 omits the control desk and permits the control operations to be initiated at the diagram board. With such an arrangement the escutcheons are located in the mimic bus in the place of symbols of circuit breakers and other devices. Two typical examples are shown in Figs. 6 and 7. The control panels of Fig. 6 are located in the same room with the power station control switchboard and are used to control the system shown in Fig. 1. This control board controls more than 600 functions at the substations within the works area, and is unique in that some of the functions are performed in motor rooms not included in the layout of Fig. 1. Fig. 7 is a view of a similar control board which has the escutcheons located in the mimic bus layout. The left-hand section of the control board controls a 25-cycle system and has the indicating instruments mounted in the mimic bus. The right-hand section controls a 60-cycle system and has the indicating instruments mounted above the mimic bus layout.

Remote Substations

It is customary in most new installations to include the remote control system as an integral part of the local control equipment in the remote substations. In general, the control wiring and instrument circuits are simplified if the local control and remote control equipments are made as a single unit. When remote control is added to an existing substation, the remote control equipment must be furnished as a separate unit.

Fig. 8 shows the local and remote control equipment as an integral unit. In this view the supervisory control system is shown mounted on the right-hand section of the duplex board. The rear panel of this same section mounts the interposing relays and telemetering transmitters. The three panels to the left of the supervisory control equipment mount the protective relays and auxiliary power distribution circuits. The rear panels of these three sections mount the control switches and instruments associated with local control.

Fig. 9 is a typical example of the supervisory control as a separate physical unit. The telephone-type relays of the supervisory control system are located on the front panel as shown. The interposing relays which have heavier contacts for connection to the circuit breaker and other control circuits, and which are under the control of the telephone-type relays shown, are mounted on the rear panel of the cabinet.

Conclusions

Because of the concentration of substations in a relatively small area, telephone-type cable circuits are usually employed to provide the necessary channels for centralized control. Other conductors in the same cables are often used for any necessary continuous telemetering and for pilot-wire protective relaying. Most centralized control systems for steel works employ supervisory control.

For efficiency operation, synchronizing at some or all of the remotely controlled substations is often desirable. It has been customary to perform this synchronizing manually rather than by automatic means. The necessary synchronism indication for manual synchronizing has been provided by connecting the necessary potentials to a standard synchroscope as well as by the use of a synchronism telemetering transmitter and an associated milliammeter.

Telemetering is an integral part of most power system centralized control installations. Most telemetering is provided on a selective or on-call basis, although continuous recording of power interchange with electrical utility companies is usually continuously recorded.

A diagram board which shows the power system circuits and which continuously indicates the position of all remotely controlled circuit breakers and motor-operated switches, has been provided in the control center of most large centralized control installations. In some installations the control functions are performed from a control desk, while in others, they are performed at the diagram board.

References

1. Supervisory Control for the World's Most Modern Steel Mill, P. B. Garrett, M. E. Reagan. *AIEE Transactions*, vol. 63, May 1944, pp. 259–64.

2. A-C Power System at the Fairless Works, S. S. Watkins, W. A. Derr, L. L. Fountain, R. B. Squires. *Ibid.*, vol. 73, pt. II, Nov. 1954, pp. 343–52.

3. Control Circuitry for Remotely Operated Electric Utility Substations, W. A. Derr, W. L. Metz. *Ibid.*, vol. 74, pt. III, June 1955, pp. 450–60.

Discussion

J. F. Headlee (United States Steel Corporation, Morrisville, Pa.): The authors should be proud of their description of a centralized control system which has found increasing acceptance in the power industry and has been used successfully in the United States Steel Corporation's Fairless Works since May 1953 and in the Geneva Works since July 1943. Based upon our operating experience at Fairless Works, we are able to comment on and emphasize several points regarding the desirability of several elements to be incorporated in supervisory control.

The first point concerns the possibility of telephone communication in conjunction with this control. Power system operators will generally agree that in emergency situations, voice communication with substations is absolutely essential. Unfortunately, the Fairless Works system did not include telephones as part of the original installation. The need for communications was dramatically demonstrated when a power outage was prolonged for approximately 45 minutes while "trouble shooters" investigated different substations and reported

their findings personally rather than by telephone.

Improvement has been obtained by adding telephone channels at the five primary substations as shown in Fig. 2 of the paper, but this has not been entirely satisfactory because it does not provide the standard of communication desired; i.e., the distribution (13.8-kv) stations should be included. Because the present telephones were added as an afterthought, the communications over these wires are interrupted by an operation and so do not allow the power dispatcher continued uninterrupted conversation. It is planned, however, to add further to this system and to include not only the secondary distribution substations, but also the yards of the primary stations as well as their control houses. We feel, however, that these comments emphasize the point made in the paper in regard to the desirability of telephone communication and the wisdom of incorporating it in the originally designed system.

With reference to system supervision by indicating lights and the use of permissive switches, our system provides such supervision by lights on the desk and mapboard shown in Fig. 4. Disagreement is shown by a flashing light and a bell. A permissive switch in each substation transfers the control from local (within the substation) to supervisory (on the desk) and the status of this control is indicated by the light on the desk.

These permissive switches are normally locked in supervisory position; the key is "captive" and cannot be taken from the lock. Since all of the switches operate from the same key as do the doors and padlocks on the power system, this captive feature provides a foreman with a reminder, in addition to an indicating light, that he has a station on local. The geographic layout of the primary power stations is such that the circuit breakers and units to be controlled are relatively close together so that there is no particular operating advantage to the use of separate permissive switches for each circuit breaker within any substation. For this operating reason and because of the economy and complications to the control circuit that would be involved, our system has one permissive switch for each substation.

As given in the paper, verification of synchronism of any of the oil circuit breakers in the remote stations in the system is required. This is given to the operator by voltmeters and a synchronoscope on the mapboard directly in front of the control desk. The method of obtaining this information is somewhat noval and ingenious. Actually, only two sources of potential are used for the entire 69-kv system; one of these is a potential transformer at the Biles Island Substation and the other is a potential transformer from the powerhouse generator bus. These two potentials are compared when any of the intermediate oil circuit breakers are being closed, and the closing circuits are so arranged that unless the synchronism has been checked the breakers cannot be checked.

Academically, this is not quite correct since it gives the curious indication that a circuit breaker has potential on one side when the circumstances obviously indicate that it does not; for example, having consecutive breakers open on the same line when reclosing any of the breakers would show potential on both sides. Despite this academic error, however, if the systems actually are out of synchronism they will always show as such. Thus, the synchronizing system accomplishes its fundamental purpose which is to assure proper synchronism and it also requires the operators to check this condition thoroughly before closing any circuit breaker where there could be a possible error.

The feature of being able to trip a circuit breaker at any local station despite the position of the permissive switch is considered extremely desirable from a safety standpoint and is being continued in the Fairless Works system despite the fact that at very infrequent intervals it boomerangs and confronts the operator with a circuit breaker tripped accidentally by unauthorized personnel.

The feature of tap changing by supervisory control is included in the Fairless Works system and allows the operators to change the taps at the interconnection transformers (those at Biles Island Substation in Fig. 2) to provide reactive power flow control and to operate the system within power factor requirements given in our power contract. This adjustment is made on the basis of reactive power flow as indicated on a meter in the control room. When this flow is excessive in either direction for the amount of real power being purchased and the required power factor provision, a tap is changed. The effectiveness of this method of operation is best shown by the fact that power factor measurements made by the utility company under their supervision have shown that our mill consistently operates at a power factor of 99.9 per cent.

Our experience and satisfaction with this equipment is best indicated by the fact that we plan to increase and expand its use to include control of air circuit breakers in the distribution substations as well as in the control of several mill motors, and in the telemetering of compressed air and fuel information through the supervisory cables to the operators at the control desk in the powerhouse.

H. G. Frostick (United States Steel Corporation, Chicago, Ill.): Centralized control of a steel-mill power system can be readily accomplished when the system is designed for such control. Steel mills having initially designed control systems are well illustrated by the authors' reference to Geneva Works and Fairless Works of the United States Steel Corporation. Undoubtedly both plants are excellent examples. Both were designed in recent years when the wide usage of electric equipment is accepted practice. Both have power systems which incorporated the newest apparatus available at the time of construction and both plants have expanded beyond their original sizes but the growth has been an orderly appended part of the parent body.

This, however, does not apply to most steel mills. There are many operating mills whose initial construction was started prior to the time that electric energy was widely used. Many plants can trace the first electric generator to reciprocating steam engines, the steam for which was obtained from the exhaust of other steam engines driving blooming and slabbing mills. Power systems in these plants reflect the sometimes haphazard results which come of expansions far in excess of original expectations. These plants, as well as some of the newest, have heavily coppered low-voltage d-c power supplies; 25-cycle services for both distribution and direct usage; and 60-cycle supplies at every conceivable voltage level up to 138 kv. Frequently, the nature of the equipment assembled in continuously expanding plants does not lend itself to co-ordinated control, but because power costs have become such a large item some control must be exercised. Centralization of this control can often be as haphazard as the growth pattern.

Steel mills traditionally generate about half or more of their power requirements. The large amounts of otherwise waste heat and by-product fuel make this economical. The in-plant generation adds to the complexity of a centralized control system because in addition to regulation of power flow, reactive flow, and voltage level, it imposes on the control system the job of regulating for the most economical usage of available fuels. In-plant generators may be gas or steam, engine-or turbine-driven units. Fuels may be blast furnace or coke oven by-products or stand-by, coal, oil, or natural gas.

Thus, where centralized control is not integrally designed into a steel mill electric system, two characteristics of the system make such control difficult to accomplish. These are the vast growth which has characterized most mills and the in-plant generation.

Another feature at Geneva and Fairless, while not unusual, makes centralized control of their power systems somewhat easier than is generally found. Geneva and Fairless are integrated mills, producing iron, converting it to steel, and in turn, manufacturing a finished product such as plate, sheet, or structurals. However, each in itself is an isolated plant not associated with other adjacent plants. More often a steel-producing activity is a complex of plants interconnected in the electric system as well as in the manufacturing process. These plants will interchange in-plant and reactive power to minimize purchased power and to obtain best efficiency from various fuels. The individual growth of each plant compounded in a complex of such interconnection arrangements does not simplify centralized control, yet it is in these plants that centralized control can be most effective.

A great deal remains to be done in this field. The apparatus available for system control in new plants generally needs modification when it is applied to older, or to combined, plants. Control is now accomplished by adaptation to local conditions. Until something more comprehensive is available, centralized control in older plants and in interconnecting plants will be as diverse and as varied as the various combinations.

W. A. Derr and W. L. Metz: Mr. Headlee has long been an advocate of centralized control of steel works power systems. He has been involved in many important installations of centralized control and has had much to do with the practices now employed for such control. We sincerely appreciate

his comments on the Fairless Works installation.

Mr. Frostick points out that it is considerably more difficult to provide centralized control for existing mills than it is to provide such control for a new and integrated works. We concur with his conclusion as to why this is true. We also concur with Mr. Frostick's statement that centralized control of existing mills can be most effective even though the application is complex. We greatly appreciate his comments.

Electric Power Systems for Steel Plants

H. N. COX
MEMBER AIEE

L. G. LEVOY
FELLOW AIEE

THE STEEL INDUSTRY is one of the largest industrial users of electricity in the United States. The electrical problems encountered include generation, transmission, distribution, and utilization.

The annual production of steel in the United States in 1955 was 116 million ingot tons. The electric energy used by the steel industry during this same year was 27 billion kilowatt-hours (kw-hr). This corresponds to about 232 kw-hr per ingot ton of production.

The nature of the steel-making process is such that a high degree of electrical service reliability is required in some areas to avoid consequential damage. In all areas, loss of production caused by electrical outages makes it necessary that a high degree of service reliability be designed into the system.

Plant Layout

Individual plants vary widely depending upon the nature of the terrain and the historical growth pattern. The plant layout, load magnitude, and distribution are important factors in the selection of the proper voltages and in the design of the electric distribution system. The nature of some of these factors can best be illustrated by an example. Fig. 1 is a layout of a hypothetical integrated steel mill which will be used for discussion purposes. This mill is located near a river or lake which provides low-cost transportation for raw materials and adequate water supplies for cooling furnaces, condensers in the powerhouse, and other miscellaneous uses. The powerhouse is located near the blast furnaces to reduce the distance the blast furnace gas must be transported to provide fuel for the powerhouse boilers, and to reduce the distance the steam for the blast furnace turboblowers must be piped.

The flow of material is as follows:

1. Raw materials are loaded into the blast furnace.
2. Pig iron from the blast furnaces, and scrap are put into the open hearth furnaces.
3. Steel from the open hearth is poured into ingot moulds.
4. Ingots are transported to the soaking pits to provide proper and uniform temperature of the ingots for rolling.
5. Ingots from the soaking pits are fed into the mills.

Coke is obtained for the blast furnace by heating coal to a high temperature for a few hours in coke ovens to drive off the tars and high Btu content gases. These gases are frequently mixed with blast furnace gas and piped to the open hearth furnaces and soaking pits for use as fuel. The residue of the coal, after the gases and tars are driven off, is almost pure carbon and is called coke.

Raw material for the arc furnaces consists of scrap and refining materials. Some plants supply molten pig iron to the arc furnaces and add scrap and other materials required to make the desired grade of steel. This is known as a hot metal charge. The power supply for the arc furnaces is usually segregated from the other loads to minimize voltage flicker problems.

Electric Energy Generation

It is theoretically possible for a well-balanced integrated steel mill with modern facilities and efficient practices to produce all of the electric power required for the various processes from the waste heat and by-product fuels available from the process.

Coal is the primary fuel. Blast furnace gas, coke oven gas, and waste heat boilers on the open hearth furnaces are the principal sources of by-product fuel and heat energy, which are utilized. The blast furnace requires more fuel than any of the other process units in the plant, and provides the largest surplus of by-product fuel, which can be used for power generation.

A blast furnace can be expected to produce about 120,000 cubic feet of gas per ton of pig iron produced. There will be some variation of this figure depending upon the quantity of coke used per unit of output. The gas is lean, averaging about 90 Btu per cubic foot, consequently it has a large volume per unit heat content. For this reason it is desirable to use it close to the source.

From 25 to 30% (per cent) of this gas will normally be used in the stoves for preheating the air to the blast furnaces. Another 25 to 30% will be used for other purposes, principally furnace blowing, soaking pits, reheating furnaces, losses, etc., leaving from 40 to 50% for the generation of electric power. Assuming a 25% station heat rate, this indicates that for each ton of pig iron produced, from 316 to 394 kw-hr of electricity could be produced. Since each ton of ingot normally contains about 0.73 ton of pig and 0.27 ton scrap on the average, then in terms of ingot tons of steel produced the excess blast furnace gas can generate from 230 to 288 kw-hr per ingot ton.

Thus a modern blast furnace rated at 1,500 tons of pig iron per day will provide enough excess gas to generate from 475,000 to 594,000 kw-hr per day or from 19,750 to 24,700 kw.

Since this excess gas is not salable for normal commercial distribution because of the excessive volume per unit of heat content, it provides a very attractive low-cost fuel for the generation of electric power by the steel plant.

Electric Energy Consumption

The amount of electric energy required to produce a ton of finished steel varies widely and depends upon whether it is produced by open hearth or electric furnaces, and upon the final finished product.

Paper 56-657, recommended by the AIEE Mining and Metal Industry and Industrial Power Systems Committees and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Summer and Pacific General Meeting, San Francisco, Calif., June 25–29, 1956. Manuscript submitted March 9, 1956; made available for printing April 23, 1956.

H. N. Cox and L. G. Levoy are with the General Electric Company, Schenectady, N. Y.

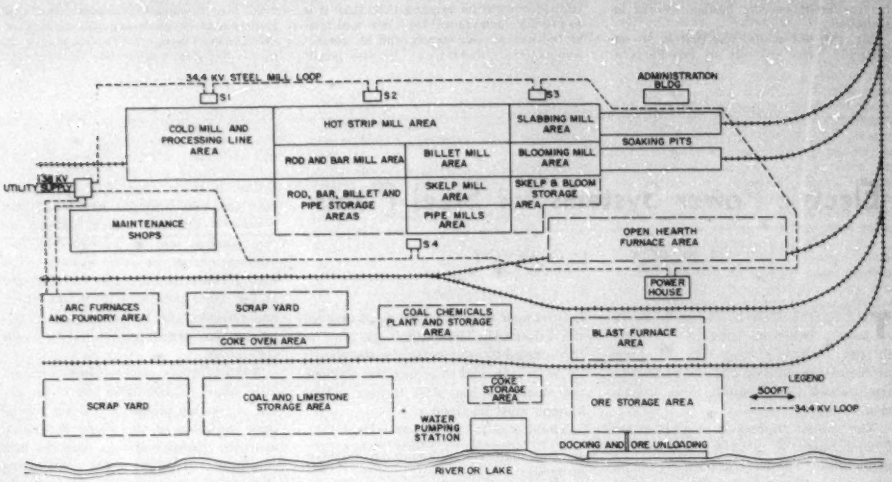


Fig. 1. Layout of an integrated steel plant

Typical figures range from 200 to 300 kw-hr per ton. Electric arc furnaces require from 500 to 600 kw-hr per ton starting with a cold scrap charge. The increasing use of arc furnaces will tend to increase the average amount of electric energy per ton of finished steel.

In an integrated plant that produces steel from raw materials in blast and open hearth furnaces (no electric arc furnaces), the power consumption will vary from 200 to 250 kw-hr per ingot ton of steel produced, when the ingots are rolled into a well-distributed list of finished products.

Table I

Process	Power Requirement per Ton of Product, Kw-Hr
Coke plant	17 (per ton of coke produced)
Blast furnaces	10 (per ton of pig iron produced)
Open hearth	8 (per ton of steel ingots)
Blooming or slabbing mill	20–25
Plate mill	30–40
Hot strip mill	80–90
Cold strip mill	80–100
Billet mill	20–30
Rod or bar mill	60–80
Structural and rail mills	30–60
Continuous pickling	6–7
Electrolytic cleaning	8–9
Tin mill annealing	15–20
Tin mill temper rolling	15–20
Electrolytic tinning	90–110
Sheet mill galvanizing	25–30

Table I lists some average power requirements for different types of rolling mills and includes all auxiliary a-c and d-c power as well as the main work roll drives. These are listed only to give a rough idea of the power requirements of different types of mills. A more accurate calculation of power requirements can be made for a given case when the particular alloy and range of products to be rolled is known.[1–2]

A plant may have more or fewer of such processes as listed in Table I, depending upon the products made. In addition to these loads, a typical plant has pumps, shops, air compressors, yard and office lighting, and other miscellaneous loads which increase the total power consumption.

The figures in Table I represent average power requirements only. Due to the nature of the various loads and their diversity, the demand will exceed the average. In a typical integrated steel plant records indicate that the 15-minute demand is 50% over the average, the 1-hour demand is 32% over the average, and the 8-hour demand is 22% over the average.

Table II indicates the range of main-drive horsepower requirements which are typical of those found in modern mills.

In addition to the type of mills mentioned here, there are many miscellaneous types and combinations.[1]

General Requirements of Steel Mill Power Systems

The steel industry, like most other modern industries, is unable to produce efficiently without the aid of electricity. The important thing to remember when designing the electric system for a steel plant is that a typical medium-size plant may have an annual output of 2,000,000 tons per year or about 5,500 tons per day, which is worth approximately $550,000 per day. Obviously, the loss of even one hour of production is costly and to be avoided if at all possible. The cost of electric energy and the equipment that distributes this power is so small in comparison to the value of the product being produced (or not being produced because of power failure), that nearly all steel companies insist on using the most reliable electric equipment that is available. Efficiency is important for economy but reliability is of paramount importance.

Table II

Type	Total Main Drive Horsepower on Work Rolls
Blooming or slabbing mills	8,000–12,000 d-c
Hot strip mill roughing train	12,000–20,000 a-c
Hot strip mill finishing train	20,000–30,000 d-c
Tandem mill	15,000–25,000 d-c
Bar or rod mills	5,000–10,000 d-c

Another reason why power failures are to be avoided in so far as possible is the possibility of contingent damage to equipment resulting from a nonscheduled power loss. The cooling water supply to a blast furnace is an example of a service which requires the utmost in service reliability. Loss of cooling water for even a short time can result in loss of the furnace and extremely costly repairs. Consequently, such services should have duplication of equipment from the water pumps back through the system to the primary sources of power. Hot metal or ladle cranes and blast-furnace skip-hoist motor-generator sets are other examples of loads that should be supplied with two or more independent sources of power. This principle of duplication of power channels should be used in all areas where a high degree of service continuity is required.

The dirty atmosphere around many steel mills results in another consideration in the electric power system design; that is, where insulators are exposed to the dirty atmosphere, high-creepage insulators are used. Examples of this are extra disks in suspension insulators, and additional creepage length on outdoor transformer and switchgear bushings. This extra creepage path length does not eliminate the need for a regular cleaning program but does lengthen the time interval between required cleanings.

Inside the mill buildings, motor rooms are provided which house the main motors' switchgear and control. These rooms are supplied with clean, filtered, and cooled air to provide a suitable environment for this equipment. The motor room in a modern cold strip mill, displaying the main drive motors and control, is shown in Fig. 2.

Fig. 2. Motor room in modern cold strip mill

Description of a Typical Electric System for a Steel Plant

Many of the design principles and techniques can best be illustrated by an example. Fig. 3 is a 1-line diagram of the primary power system for the mill layout shown in Fig. 1, and illustrates one way in which such a plant could be served.

In this arrangement, the utility is assumed to supply two 138-kv incoming lines. The two steam turbine generators shown for the powerhouse supply part of the electric power for the mill. Blast furnace gas is used as a fuel for the boilers which supply steam for the turbine generators.

A 34.4-kv primary loop circuit is used for illustration purposes in Fig. 3 to distribute power to the stepdown substations in the mill areas. Power is fed into this loop from the utility on one end and by the two plant generators on the other end, thus providing four sources of power for the loop. Reserve capacity is provided so that loss of any one generator, transformer, or primary circuit would not limit production. The 34.4-kv loop primary provides a high degree of service reliability to all substations. Load tap changing is provided on the utility tie circuit transformers to maintain essentially constant voltage on the 34.4-kv system and to permit proper division of reactive kilovolt-amperes between plant generators and utility. The transformers are connected delta-wye to provide for grounding on the 34.4-kv side.

The selection of voltage and the circuitry for the primary power distribution is frequently dictated, at least in part, by the historical growth pattern of the plant. It is possible to serve plants of this size, or even larger, at 6.9 or 13.8 kv without the use of higher voltage subtransmission circuits, but such systems do not usually provide the economy and flexibility for future growth that can be realized by use of a higher voltage subtransmission system.

The voltage of the subtransmission circuit should be selected in any given case at the level which provides for economic system design. Transmission circuits and associated equipment should have adequate reserve for emergency operation as required, and the system should be flexible for future growth. A higher subtransmission voltage such as 69 kv would probably be used if the plant were larger and so required longer transmission lines and heavier loads. In large plants, multiple subtransmission circuits are frequently used to provide improved reserve capacity and service reliability.

In situations where the utility requires that the 138-kv winding be grounded, wye-delta-wye transformers can be used to permit grounding on both sides. Lightning arresters, relays, and surge protective equipment have been omitted from this diagram for simplification.

It is possible to substitute horn-gap disconnect switches for the 138-kv transformer primary breakers and to interlock these switches with the secondary breakers so that they will be used only to open transformer magnetizing current. One disadvantage of this arrangement is that in event of a transformer fault, it is necessary to trip one 138-kv incoming line and the bus tie breaker, thereby deenergizing both transformers on one bus until the necessary switching can be accomplished to restore the sound transformer to service.

In the normal operation of the plant shown in Fig. 3, the arc furnace load is segregated from the other mill loads by use of separate 138–34.4-kv transformers. This arrangement minimizes the voltage flicker disturbances on the circuits feeding the other mill loads. A normally open tie breaker is provided between the arc furnace and mill loop supply busses at the 34.4-kv level for flexibility of operation in event of a transformer outage.

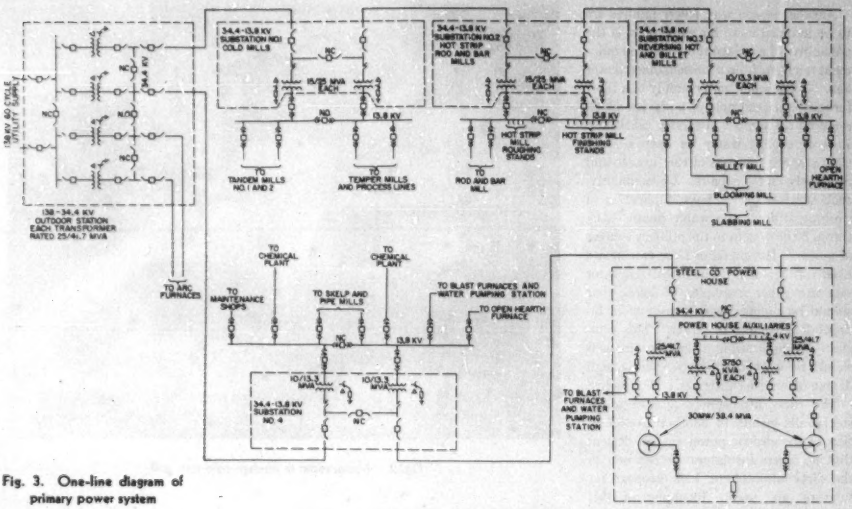

138 KV 60 CYCLE UTILITY SUPPLY
138-34.4 KV OUTDOOR STATION EACH TRANSFORMER RATED 25/46.7 MVA
TO ARC FURNACES
34.4-13.8 KV SUBSTATION NO.1 COLD MILLS
15/25 MVA EACH
TO TANDEM MILLS NO. 1 AND 2
TO TEMPER MILLS AND PROCESS LINES
34.4-13.8 KV SUBSTATION NO.2 HOT STRIP ROD AND BAR MILLS
15/25 MVA EACH
HOT STRIP MILL ROUGHING STANDS
HOT STRIP MILL FINISHING STANDS
TO ROD AND BAR MILL
34.4-13.8 KV SUBSTATION NO.3 REVERSING HOT AND BILLET MILLS
10/13.3 MVA EACH
BILLET MILL
BLOOMING MILL
SLABBING MILL
TO OPEN HEARTH FURNACE
TO CHEMICAL PLANT
TO CHEMICAL PLANT
TO MAINTENANCE SHOPS
TO SKELP AND PIPE MILLS
TO BLAST FURNACES AND WATER PUMPING STATION
TO OPEN HEARTH FURNACE
10/13.3 MVA
10/13.3 MVA
34.4-13.8 KV SUBSTATION NO. 4
STEEL CO. POWER HOUSE
34.4 KV
POWER HOUSE AUXILIARIES
2.4 KV
25/46.7 MVA
25/46.7 MVA
2.4 KV
37/62 MVA EACH
13.8 KV
TO BLAST FURNACES AND WATER PUMPING STATION
30MW/38.4 MVA

Fig. 3. One-line diagram of primary power system

Each of the four substations is provided with 34.4–13.8-kv stepdown substations to provide power at 13.8 kv to the various mill areas. The 34.4-kv tie normally is operated closed at each substation to provide dual channels for power flow to all stations under normal operating conditions. All 13.8-kv busses (except the power station bus) in this system are laid out for use with standard metal-clad switchgear of 500-mva (megavolt-ampere) interrupting capacity. A bus tie circuit breaker is provided in each 13.8-kv substation bus. It is desirable to operate with this circuit breaker closed in cases where closing it will not cause the fault levels to exceed the 13.8-kv switchgear rating. Whether or not this circuit breaker is operated normally closed depends upon the size and impedance of the supply transformers and the characteristics of the connected load.

For example, in Fig. 3 the 13.8-kv bus tie circuit breaker in substation no. 1 is shown normally open because these mills are normally driven by d-c motors supplied from synchronous-motor-driven motor-generator sets which will contribute substantially to the fault duty on these busses. However, in substation no. 2 the 13.8-kv bus tie is shown normally closed even though the same size transformers are used. In this case the hot strip mill finishing stands are driven by d-c motors supplied from power rectifiers which do not contribute to the fault duty on the bus, thereby making it possible to operate with the bus tie closed, even though the roughing stands and bar mill motors do contribute to the fault duty.

Recently, 13.8-kv metal-clad switchgear rated at 750 mva has been developed. This will permit the design of systems with higher fault levels in cases where this is warranted, and yet maintain the well-known advantages of metal-clad switchgear.

Note that two feeders are used to supply each important load. In cases where the bus sections are operated with normally open ties the two feeders to the subdistribution bus in a given load area originate from the same bus section. This eliminates the possibility of accidentally connecting the busses together at the subdistribution bus with the consequent increase in fault levels beyond the circuit breaker ratings. In those substations which can be operated with the 13.8-kv bus tie circuit breaker normally closed, duplicate feeders to a given load area originate from different bus sections to improve the service reliability.

In the powerhouse, cubicle-type switchgear of higher interrupting rating is used on the main 13.8-kv busses because the fault duty on these busses exceeds available metal-clad switchgear ratings. The power plant auxiliaries are supplied by means of stepdown transformers as indicated. For power which is to be used nearby at 13.8 kv, such as the feeder shown for the blast furnace and water treatment plant, a current-limiting reactor is used to reduce the fault duty so that 500-mva metal-clad switchgear can be used in such areas. For larger power plants containing more generating units, a synchronizing bus arrangement may be used to limit the fault duty.

Fig. 4 shows one possible arrangement of the power distribution circuits for the area supplied by substation no. 2. Although the 13.8-kv bus tie can be operated closed in this case, the loads have been arranged so that it can be operated open if desired. This provides added flexibility for additions and changes which may be required from time to time on various busses or circuits.

Note that the finishing stands are all connected to the same a-c bus section. This is desirable because the six large d-c motors receiving power from the rectifiers must have a constant-speed relationship with each other. If different a-c sources were used to supply these rectifiers, a momentary variation in voltage on one source might disturb the speed relationships momentarily until corrected by the control. Also, harmonic cancellation may not be as effective.

Note also that four phase-shifting transformers are used with the six rectifier transformers. This provides some harmonic cancellation among the rectifiers so that the finishing stand load appears on the a-c bus as a 36-phase recti-

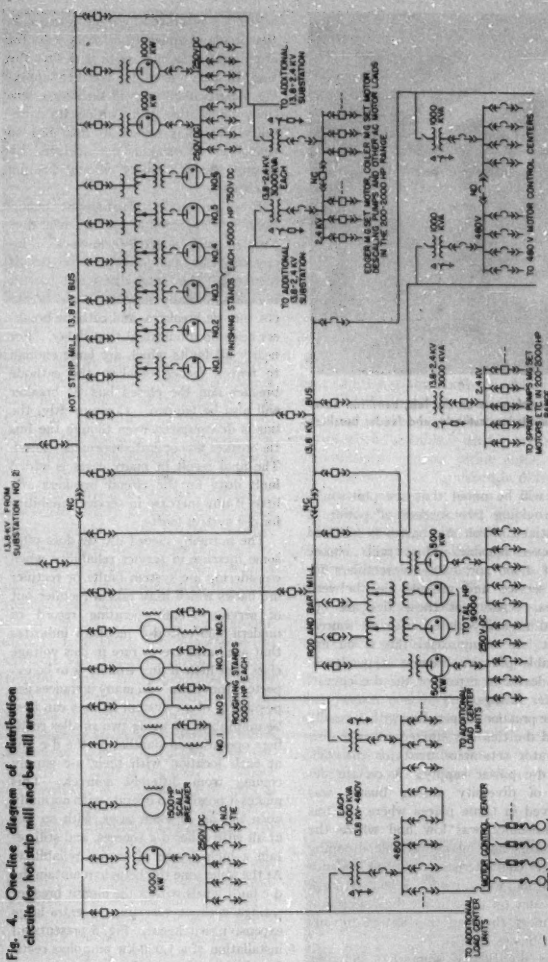


Fig. 4. One-line diagram of distribution circuits for hot strip mill and bar mill areas

fier load with a resultant low harmonic content. Either load or no-load tap changing can be provided to give economical, high-power-factor operation of the rectifier at reduced voltage. If a major adjustment of direct voltage is required while the mill is running, step-voltage-regulating transformers can be used. If the mill can be shut down when a major change in direct voltage is required, no-load (de-energized) tap changers on the rectifier transformer may be used.

Each rectifier unit is equipped with anode breakers to clear an arc back rapidly with a minimum of disturbance to the mill or system. Automatic reclosing of these anode breakers after a time delay is provided.

In planning the power supply for rectifiers, it is desirable to minimize the a-c regulation and voltage spread to decrease the amount of phase control required, since rectifier power factor varies with the amount of phase control used.

The end of the mill into which the slab first enters from the slab reheat furnaces is called the "roughing end." At this end of the mill the slab is short and relatively thick, and so the stands are spaced so that the slab is in only one stand at a time. Many of the older hot strip mills started with a narrow slab cut to a length equal to the first stand roll length. The first reduction would occur in this stand when a "broadside pass" was made to widen the slab to the desired width of the final strip. A wound-rotor induction motor with flywheel was used to drive this first stand, because the broadside pass would last for only a couple of seconds. The short time involved made it feasible to make use of the stored energy in the motor rotor and flywheel, and in this way a smaller horsepower motor could be used. Modern hot strip mills start with a slab as wide as is required and up to 20 feet in length. With such a slab the steel is in each roughing stand for 5 or more seconds. This longer time makes it impractical to use a flywheel and wound-rotor motor; consequently, synchronous motors are being used in the newer mills. The wound-rotor motors are started at full voltage with resistance introduced in the rotor circuit by means of the slip regulator. The large synchronous motors may be started on full voltage or on reduced voltage, usually with neutral reactor start.

Large synchronous motors which are subjected to rapidly varying mechanical loads are usually supplied with power factor regulators to minimize the var swings and resulting voltage swings which would otherwise appear on the system.

Both the 2,400 and 480-volt services are derived from the 13.8-kv by a single transformation, usually using a secondary selective arrangement with the primaries of each pair connected to opposite sides of the 13.8-kv bus for maximum flexibility and continuity of service. The substation sizes are selected for economy of switchgear and to keep the secondary cable runs as short as is practical. This minimizes installation cost, secondary losses, and voltage regulation.

Selectivity is provided in the high- and medium-voltage systems by appropriate relaying. In the 480-volt system, selectivity is obtained where required by proper selection of low-voltage circuit breakers and control. In any set of auxiliaries where the process can continue with one auxiliary circuit out of action, or where the unscheduled shutdown of a group of auxilaries will cause severe consequential damage, it is desirable to provide for selectivity.

Fig. 5. 1,000-kw pumpless rectifier showing rectifier transformers on the left, transition compartment, anode breakers, pumpless rectifier, rectifier auxiliaries, cathode, and feeder breakers

Load center unit substations and motor control centers are almost universally accepted in the steel industry for auxiliary power and control. Combination starters are available with either fuses or circuit breakers for fault protection.

D-C Systems

Main drive motors are usually designed for 600-, 700-, or 750-volt operation. It is normal practice to operate these d-c systems ungrounded. On large common bus systems fault levels tend to become excessive; consequently many modern, large, high-production mills use the individual generator or rectifier system. While this results in a lack of diversity on the sources so that full source capacity is required, it does not necessarily mean that any less would have been provided on a common bus. The use of the individual generator or rectifier system results in a substantial saving in low-voltage bus and switchgear, and also brings the d-c fault levels down to acceptable values.

Steel mills usually require a reliable source of 250-volt constant potential direct current for the operation of cranes, screwdowns, coil-handling equipment, etc. Prior to 1940 this power was usually furnished by motor-generator sets. Improvement in design of rectifiers has made them a reliable source of power for this service and they are being used in most of the recent installations. Some of the advantages of rectifiers as compared with motor-generator sets for this service are lower installation cost, less maintenance, and higher efficiency.

It will be noted that the philosophy of providing two sources of power to essential 250-volt d-c busses is followed wherever feasible. Most mills choose either 1,000- or 1,500-kw rectifiers for this service and provide ties between busses. Sometimes these ties are operated normally closed with all sources intact. In many cases this is an undesirable practice because it results in fault levels in excess of the d-c circuit-breaker ratings.

The practice of operating with normally closed d-c ties was started when motor-generator sets were used for the 250-volt d-c power supply. A certain degree of diversity among busses was achieved in those places where the bus tie resistance was low and where the generators had an appreciable drooping voltage characteristic, or load balance regulators were used. When a large load came on a particular bus, a certain portion of the load was shared by adjacent busses.

The situation is somewhat different when rectifiers are used as the source of 250-volt d-c service. The rectifier is capable of rapidly assuming additional load and its 250-volt bus is regulated close to 250 volts. Load balance circuits may be used but since most steel mill engineers like to have the 250-volt bus voltage relatively steady, such balancing circuits are designed to be relatively weak. With rated alternating voltage applied, the rectifier will assume up to 150% of its rated load and still maintain essentially 250-volt direct current at the bus. In many cases the d-c busses are separated by some distance so that it may require 10 or more volts for any appreciable current to flow over the tie. In such a case, the rectifier supplying a particular bus will pick up essentially all of the load (within its capability) which is added to that bus so that load diversity is not achieved by closing the ties under normal loading conditions.

However, under short-circuit conditions, an appreciable current will flow through such normally closed ties, increasing the fault duty on the circuit breakers. The semi-high-speed circuit breakers normally supplied as d-c breakers, bus tie breakers, and cathode breakers have no intentional time delay. For nearly all faults which are large enough to trip a feeder breaker, the cathode breaker and the closed bus tie breaker will also be tripped. Consequently, the bus is de-energized even though the bus tie breaker was operated normally closed. The final result in many cases is added fault duty on the circuit breakers and little if any increase in service reliability for d-c system faults.

The normally closed d-c tie does offer some increase in service reliability when considering a-c system faults or rectifier arc backs which may take a rectifier out of service. The operating record of modern 250-volt d-c rectifiers indicates that arc backs are so rare in this voltage class that little if any trouble is to be expected from them. In many instances the possibility of a-c system failures can best be minimized by using two smaller rectifiers operating in parallel on the d-c bus at each location with their a-c supply coming from different sources. This makes it possible to operate with normally open ties or, in some cases, with no ties at all with other d-c sources, and still retain a high degree of service reliability. At the same time this helps to maintain the d-c fault levels within the circuit breaker ratings without resorting to extra-large expensive switchgear. Fig. 5 presents an installation of a 1,000-kw pumpless rectifier showing the rectifier transformer on the left, transition compartment, anode breakers, pumpless rectifier, rectifier auxiliaries, cathode breaker, and feeder breakers.

Cable Selection

As with all other equipment in the steel industry which can affect production, cable must be characterized by utmost reliability. For many years lead-jacketed cable was used almost exclusively in steel plants for the distribution of power in the

Fig. 6. Conduit in steel mill motor room floor prior to pouring concrete

2.4- to 13.8-kv voltage range. However, within the last decade many steel plants have been using the newer types of cable which are insulated with synthetic rubber-like compounds. The advantage of these cables is the elimination of potheads which are essential on lead-jacketed cables. Potheads are expensive to install and require a high degree of skill to make them as reliable as the rest of the cable. The building of stress cones and the taping of the synthetic rubber-like insulated cables at the termination does, of course, require some skill.

For the 2.4- to 13.8-kv underground circuits, nonmetallic conduits encased in concrete are frequently used. In each conduit, three single-conductor cables with ozone-resistant synthetic-rubber insulation and neoprene jacket are used. For voltages above 2 kv, shielded cables are recommended. With this cable system, no pothead is required for indoor terminations. For outdoor terminations where no larger terminal box is available, a pothead with stuffing box in place of the wiping sleeve can be used.

In some instances 3-conductor varnished-cambric-insulated shielded lead-jacketed cable with an outer covering of a neoprene compound is used in place of the three single-conductor cables. This cable requires a pothead and is particularly applicable where cable connections to an existing lead-covered cable system are required or in places where severe electrolytic corrosion conditions would shorten the life of ordinary lead-covered cable.

For circuits above ground which run under the main motor room floor beams or through roof trusses, 3-conductor varnished-cambric-insulated shielded cable with a thermoplastic sheath and interlocked armor cover is commonly used. These cables are supported by ladder-type racks and no potheads are required for indoor terminations in dry locations. A recent addition to this family of cables is an extruded thermoplastic coating over the armor to protect against corrosion of the armor when the cable is used in corrosive atmospheres such as pickle lines, etc. The coating can be furnished in different colors as a convenient means of identifying various voltage levels, which saves painting or taping of the cables for identification. Synthetic-rubber-type insulated cables can also be used in these locations.

For 600 volts and below, single-conductor rubber-insulated type-*RH-RW* synthetic-jacketed cables are commonly used either in trays or conduits. Fig. 6 shows an installation of conduit in a motor room floor prior to pouring the concrete. Fig. 7 shows a tray-type installation in a motor room basement. This is used instead of conduit and is an alternate to the construction shown in Fig. 6. Heavy-duty circuits, such as power feeds from substations to load centers which are run above ground, are often served with multiconductor varnished-cambric-insulated cable with a thermoplastic sheath covered with interlocked armor. Three-conductor cables of this type are normally used for a-c power and 4-conductor cable has been found to be advantageous for d-c circuits, using two conductors in parallel for the positive and the other two for the negative side of the circuit.

The 600-volt d-c circuits may be made with single-conductor varnished-cambric-insulated cable with a flame- and moisture-resistant braid finish in dry locations or with single-conductor type-*RH-RW* synthetic-jacketed cable.

Interconnecting control cable running between equipments usually has a minimum size of no. *12* American Wire Gauge stranded. Multiconductor cables are generally used for control circuits. The individual conductors are covered with colored thermoplastic insulation to provide a means of color coding. A glass-reinforced paper tape is used to bind the individual conductors in a cable and a thermoplastic jacket is placed over all.

In the selection of all cables, it is important not only that the cables have adequate current-carrying capacity but, in addition, voltage-drop limitations and fault-duty requirements should be recognized. The fault-duty limitation on the minimum size of any cable is established by the available short-circuit current which must not give rise to excessive temperature in the cable during the time it takes the protective means to open the circuit under fault conditions.

Fig. 7. Tray-type installation used instead of conduit for low-voltage power and control cables

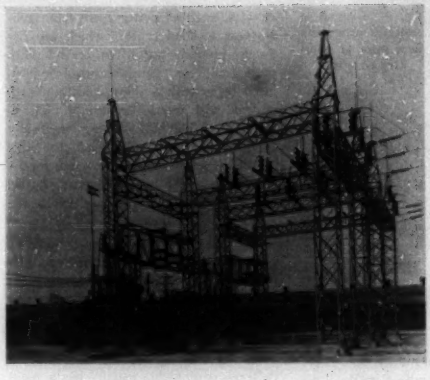


Fig. 8. Outdoor package substation for a steel mill showing incoming lines, high-voltage circuit breakers, power transformers, lightning arresters, disconnects, steel structure, and control house which contains the instrumentation, control equipment, and station battery

Trends in Steel Plant Electric Systems and Equipment

There appear to be certain trends which are discernible in steel mill electric systems and equipment as these systems evolve through experience and availability of improved designs. Some of these are discussed in the following.

Generated Versus Purchased Power

The trend today, which is the result of many economic studies made by various plants, is to utilize the waste heat and by-product fuel available from the process to generate power. This is supplemented by coal or oil in periods when a sufficient amount of by-product fuel is not available to utilize the installed generating equipment. For power requirements beyond this, the trend is to purchase the additional power requirements from the electrical utility system. This arrangement makes available capital for the purchase of new efficient steel production or rolling equipment, or for plant expansion which otherwise would have to be used for the purchase of power-generating equipment.

A natural result of the practice of both generating and purchasing power is a tie between the steel plant system and the electrical utility system. This tie brings about certain problems such as the proper sharing of kilowatt and kilovars between the two systems. A discussion of these problems has been presented elsewhere.[4,5]

Selection of Voltages

It is important to select the proper voltages to serve the loads economically and provide for possible future expansion. Many of the older mills are handicapped by starting off with a voltage which was adequate at the time it was selected, but is too low for their present size.

Larger plants usually find it desirable to have a high-voltage circuit such as 34.4 or 69 kv to distribute the power to spot loads on the site and then transform down to 13.8 kv for utilization at these load centers. The circuits shown in Fig. 3 illustrate this type of system. This plant is relatively compact and it would be possible to serve it at 13.8 kv. If the plant were spread out, requiring longer transmission circuits between spot loads, 34.4 kv would show greater economic advantage; 69 kv would be economic for a larger plant requiring still more power and greater transmission distances. Because of the size of load involved, the tie with the utility system is often made at 115 or 138 kv.

While 6.9 kv has been used extensively in the past as a utilization voltage for large motors in steel mills, there is a trend toward the use of 13.8 kv because the switchgear ratings available make it possible to handle more load at 13.8 kv than at 6.9 kv without exceeding the momentary rating of the circuit breakers. Also, 13.8 kv provides greater flexibility for picking up a larger load at some distance and reduces cable costs for serving large loads. The disadvantage of 13.8 kv as a utilization voltage is that motors in the 2,000–15,000-horsepower range cost more at 13.8 kv than they do at 6.9 kv.

Economic studies in many steel mills have indicated that the following choice of motor voltages provides a flexible and economic system:

System Voltage	Motor Voltage	Motor Horsepower
13.8 kv	13.2 kv	2,000–15,000
2.4 kv	2.3 kv	200–2,000
480 volts	440 volts	200 and below

Exceptions will occur where it will be desirable to deviate from this guide, depending upon the length of circuit required to serve the motor at the preferred voltage.

Outdoor Stations

Fig. 8 shows a typical outdoor package substation for a steel mill. This station steps the voltage down from 115 to 13.8 kv. The illustration shows the incoming lines, high-voltage circuit breakers, power transformers, lightning arresters, disconnects, steel structure, and the control house which houses the control equipment and station battery.

High-voltage bushings for such transformers and switchgear are frequently specified with extra creepage length if severe dirt conditions are to be encountered. The power transformers for such a station are normally of the oil-filled type. In cases where the plant contains generation which is operated in

Fig. 9. Lineup of modern 13.8-kv metal-clad switchgear

parallel with the utility, it is desirable to provide for load tap changing control on the tie transformers for controlled division of kilovars between the utility and local generators, and at the same time maintain the plant voltage essentially constant.

Load tap changing is also desirable in situations where the plant purchases all of its power from the utility, but the utility voltage varies over too wide a range to maintain satisfactory voltage levels in the plant with a fixed transformation ratio.

It is frequently found desirable to provide additional transformer capacity for overload or emergency conditions. Generally this can be done most economically by selecting transformers with fan-cooled ratings such that the normal loads can be carried by the self-cooled ratings and the overload or emergency condition satisfied by the fan-cooled rating.

These transformers are usually specified with the 13.8-kv winding wye connected with the neutral brought out for resistance grounding. The grounding resistors in the station shown in Fig. 8 are of the outdoor type and are mounted in the station structure.

Medium-Voltage Switchgear

The 13.8-kv switchgear normally used in modern steel mills is of the metal-clad vertical lift or drawout type and is oilless, thereby eliminating the disadvantages and hazards of oil. Very often these circuit breakers are located indoors in a motor room. However, they can be provided with a weatherproof housing and located outdoors. Fig. 9 shows a lineup of modern indoor 13.8-kv metal-clad switchgear.

The requirements of 13.8-kv 500-mva switchgear for steel mill service are such that it may differ from general-purpose switchgear in the following respects:

1. The bus is normally specified for 2,000 amperes instead of a 1,200-ampere bus as is commonly used in general-purpose switchgear.

2. The circuit breakers are specified to have 60,000-ampere momentary rather than the 40,000-ampere momentary called for by the American Standards Association standard general-purpose design.

In the past, vertical life or drawout switchgear has been limited to a maximum of 500-mva interrupting capacity. A new design of metal-clad switchgear rated 750-mva interrupting capacity is now available and will undoubtedly find application in steel plants as well as in other industrial and utility plants.

The 2,400-volt switchgear in modern mills is also generally of the metal-clad vertical lift or drawout oilless type of 150-mva interrupting rating. Frequently this

Fig. 10. Double-ended load center unit substation, 2,000 kva, 13.8–480 volts, showing incoming line switches, transformers, and secondary switchgear

2,400-volt service is derived from the 13.8-kv service by use of package equipments consisting of a 13.8–2.4-kv step-down transformer, flange, or throat connected to the metal-clad switchgear. These transformers should be connected 13.8 kv delta, 2,400 volts wye, with the neutral brought out for resistance grounding. For indoor installations, askeral-filled transformers are normally used.

Low-Voltage Substations

The 480-volt service is normally derived from the 13.8-kv service by a single transformation. Here again package units are normally provided which consist of an incoming line switch, askeral-filled transformer, and drawout 480-volt air circuit breakers. Frequently, double-ended units are provided with two transformers and their associated switchgear in a secondary selective system with a normally open tie between the 480-volt transformers to provide for emergency service. This arrangement makes it possible to use economical switchgear ratings. Fig. 10 shows a 2,000-kva 13.8–480-volt double-ended substation with incoming line switches, transformers, and secondary switchgear. These transformers should be connected 13.8 kv delta, 480 volts wye, with neutral brought out for grounding.

System Complexity

Older systems were often designed with many possible routes for power to flow to a given load. This was done with the intention of improving service reliability. However, this resulted in relatively complex systems with complicated operating rules to avoid excessive fault duty on the breakers. Many accidents and failures have occurred in such systems due to errors in switching, particularly during emergencies. Simple systems minimize the possibility of such errors.

Improved apparatus and cables have eliminated the need for much of this complication, and the trend today is to obtain service reliability by the use of high-quality components carefully installed in relatively simple circuits which are easy to understand and operate.

Safety

The use of metal-clad switchgear has eliminated the need for many disconnect switches which were always a source of danger to operators. There is a definite trend away from the indiscriminate use of disconnect switches in the interest of safety and improved service reliability.

System Frequency

Some of the older steel mills use 25-cycle power on part or all of their system. There were a number of reasons for this at the time these systems were designed, such as the availability of 25-cycle power in some areas and the use of synchronous converters to supply d-c power. These converters had inherently better operating characteristics at 25 cycles. Also, better 25-cycle motors could be built at very low speed for direct connection, and gears were not as well developed as they are today. Most of the reasons for the original choice of 25 cycles have disappeared and 60-cycle systems dominate the field in the United States. Due to the better availability and lower cost of 60-cycle equipment, most mills are no longer expanding their 25-cycle systems, and any major changes or additions are usually made at 60 cycles. A 25/60-cycle frequency tie is usually provided between the 25- and 60-cycle systems in those plants which have both frequencies. This provides flexibility for the interchange of power between the two systems.[8]

Grounding Practice

The advantages of a-c system neutral grounding at all voltage levels is being recognized. Many existing plants have changed from ungrounded to grounded neutral operation[7] and most new installations are planned for neutral grounding. It will be observed from Fig. 3 that the 13.8- and 2.4-kv systems are resistance grounded to minimize possible burning damage on the motors connected to them, and yet provide selective relaying of faulted circuits. The 480-volt system is solidly grounded since at this voltage level the impedance of the ground return path is usually sufficient to limit the ground fault current to acceptable values.

Closing and Tripping Batteries

It is considered good practice in steel plants to use 250-volt d-c batteries for closing and tripping power on the mill switchgear. A voltage of 250 is used in preference to 125 volts because there is less chance of failure to operate due to dust accumulating on switch contacts, etc., especially when there are many interlocks connected in series, and is therefore considered to be more reliable. In addition, 250 volts provide less voltage drop for operating mechanism of circuit breakers located at a greater distance from the battery, and future expansion can be more easily accommodated.

Lead acid batteries with suitable chargers have been used for many years and are still being widely used, but nickel-cadmium batteries are being requested more and more because of their long life and cleanliness.

D-C Supply and Circuitry

Rectifiers are gaining acceptance for use on 250-volt constant-potential d-c systems and as a source of adjustable direct voltage for the main drive motors of continuous hot mills. Rectifiers are being used instead of motor-generator sets in new installations because of their lower installation cost, lower maintenance, higher efficiency, and faster response rates.

Steel mill engineers are becoming more and more conscious of the need for adequate switchgear on d-c systems to provide for fault protection in much the same way as adequate fault protection has been recognized on low-voltage a-c systems. This awareness is causing the segregation of d-c systems into smaller units to avoid excessive fault levels.

Conclusions

Historically, the steel industry has been quick to recognize new electrical developments and put them into use. Progress in electric systems and components is continuous. Increased production requirements and the trend toward automation will further increase the requirements for improved and expanded steel plant electric systems.

References

1. The Making, Shaping, and Treating of Steel. U. S. Steel Corporation, Pittsburgh, Pa., sixth edition, 1951.

2. Power Consumption of Hot Strip Mills. W. M. Ballenger, T. R. Rhea. *Proceedings*, Association of Iron and Steel Engineers, Pittsburgh, Pa., Mar. 1941.

3. Power Requirements and Selection of Electric Equipment for Hot Strip Mills. R. E. Marrs. *Iron and Steel Engineer*, Pittsburgh, Pa., July 1950, pp. 111–25.

4. Parallel Operation of Industrial Power Plants with Public Utilities. L. G. Levoy. *Distribution*, Schenectady, N. Y., Nov. 1953.

5. Kilovar, Kilowatt and Voltage Control in the Steel Plant. L. G. Levoy. *Proceedings*, Association of Iron and Steel Engineers, 1955, pp. 459–67.

6. Frequency Changers—Characteristics, Applications, and Economics. S. B. Crary, R. M. Easley. *AIEE Transactions*, vol. 64, June 1945, pp. 351–58.

7. Power Generation and Industrial Power Systems. *AIEE Special Publication S-38*. "The Use of 600 Volt and 460 Volt Power Systems with Grounded Neutrals," James P. E. Arberry. Aug. 1950, pp. 187–88.

8. Industrial Power Systems Handbook, D. L. Beeman, editor. McGraw-Hill Book Company, Inc., New York, N. Y., 1955.

Discussion

I. N. Tull (Republic Steel Corporation, Cleveland, Ohio): The authors show their broad and detailed knowledge of the electrical requirements of steel plants in their discussion. Such an item as proper distribution voltages is sometimes hard to predetermine properly, and there are many steel plants operating at present on voltages far too low for economical distribution and switching mainly because no one involved in the original design of such plants had any idea of the extent to which the power requirement would grow.

We agree with the rule of thumb used by the authors for the establishment of motor and plant distribution voltages. It has worked out well in several of our more recent installations.

The grounding of neutrals has improved our cable and apparatus performance; since installation on our 460-, 2,300-, and 11,000-volt distribution systems no multiple failures have occurred at times of trouble, and the consequential damage has been far less than that encountered before grounding. We do have more trouble in locating the points of failure, due to the limited damage, but that is so much better than having multiple failures, lamination burning, and more frequent failures that we do not contemplate any new installations of ungrounded systems.

The service record of 3-phase transformers over an extensive time experience has practically eliminated new installations of single-phase units that could be operated open delta on failure of one unit, so the wye grounded secondaries system is no longer considered a serious handicap. Of course, in our large plants, flexibility of supply from two or more circuits is mandatory.

Reliability of and continuity of current supply is paramount in our industry. The natural hazards we have in handling huge volumes of gas and liquid metals at high temperature with proximity of enormous quantities of water makes it imperative that lights in many locations must be depended on, as well as water and other services.

Some system designers now feel so confident that switchgear busses are indestructible, and to alleviate the high costs of sectional busses advocate a single bus. However, accidents unfortunately can happen and extensions do have to be made, as well as changes in the rapidly growing steel industry so that, for important areas, we must subdivide our busses, transformers, and lines and arrange our switching for quickly meeting situations that arise only too frequently.

Our distribution system units should not cover too great an area, particularly auxiliary d-c circuits, as it is necessary on account of the danger of electrolysis to operate such circuits ungrounded. With large numbers of motors, small and large, and their associated feeders, it is difficult to locate accidental grounds when they develop, many times multiple, unless the system can be quickly split into small isolated sections.

When large numbers of small- and medium-size motors are installed in a given area, it is a temptation to use transformers of too large a rating for the ability of the motor controls successfully to interrupt fault currents, and we restrict our units to 1,000 kva at 460 volts, with normal impedance of 5%. If too large transformers are provided, the motor controls and individual protection become too expensive.

Ward-Leonard control is essential in many cases of main-drive and specialty operations, and is desirable in a great many other cases. But where it is used for instance to supply blast furnace hoists and usually adjacent docks, the costs are sometimes prohibitive as compared to magnetic controls operating from a common d-c source.

In the latter case, when trouble develops usually there are enough spares common to the group to enable us to start quickly, but unless a Ward-Leonard control is provided with alternate equipment or the impossible economic duplication of the motor-generator sets is provided, some days might be lost in rewinding the motor or generator, and that of course is entirely unthinkable. In either case, the distribution and switching system must be entirely reliable to meet our requirements. If the general principles set

out in the paper are followed, I am sure, from our experience, that reasonable and sensible results will be obtained.

H. N. Cox and L. G. Levoy: We appreciate the comments of Mr. Tull who has been closely associated with the operating problems of steel plants for many years. We are happy to know that his experience generally confirms the ideas expressed in the paper.

With regard to the use of Ward-Leonard control on blast furnace skip hoists, for a wide speed range (about 4 to 1) the usual practice is to use two Ward-Leonard drives for the same hoist, usually series connected for equal load division. The preferred arrangement is to have the a-c power supply for the two motor-generator sets coming from a split bus or separate busses. The equipments are sized so that on loss of one unit (a-c motor, d-c generator, or hoist driving motor) the remaining unit can operate the hoist at a reduced speed, but adequate for keeping the furnace in operation. This provides for a high degree of continuity of service.

For skip hoists which operate at relatively low skip speeds and where the speed range is narrow (3 or 4 to 1), constant-potential direct current with armature series and shunting resistors may be used with field control. Similar duplication in power supply circuitry, hoist motors, and reserve capacity to provide for continuity of operation on loss of one unit is desirable.

Voltage Dips and Flicker

A. A. KRONEBERG
FELLOW AIEE

POWER SYSTEMS of electrical utilities are nominally constant-speed (mechanical) and constant-voltage (electric) types of system. Therefore, the degree to which these two quantities are maintained constant at all times determines the quality of service supplied by the utility. Changes in mechanical torques and in electric currents on the system cause deviations from normal speed and voltage. Adjustments in input to prime movers, in excitation of synchronous machines, and in ratio of transformation are required to restore balance on the system with normal speed and voltage. Large capital investments are made by the utilities for equipment designed for automatic control of speed and voltage. All such equipment has an inherent time-delay characteristic. First, a deviation from normal value must take place before the need for correction is detected by the regulating device. Then, the corrective action must be damped or delayed enough to prevent oscillations. The magnitude of the controlled quantity is held within a band between a minimum and a maximum value for relatively slow changes of loading.

Sudden changes in current or torque will cause deviation from normal in voltage or speed outside the band width of normal regulation before a correction can be made to bring it back within the band. A voltage dip is an occasional single disturbance caused by short circuits, faults, and starting inrush of current to motors. Voltage flicker is a series of recurring voltage dips. Cyclic or rapidly recurring voltage changes are generally more objectionable than noncyclic.

Paper 56-637, recommended by the AIEE Industrial Power Systems and Mining and Metal Industries Committees, and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Summer and Pacific General Meeting, San Francisco, Calif., June 25–29, 1956. Manuscript submitted March 12, 1956; made available for printing April 17, 1956.

A. A. KRONEBERG is with Southern California Edison Company, Los Angeles, Calif.

Limits of Tolerance

Voltage dips caused few complaints as late as 2 or 3 years ago. The emphasis on automation and the assumption by manufacturers of automatic machines, control contactors, and relays of an absolutely continuous power supply have resulted in increasing complaints from customers concerning troubles due to voltage dips or momentary interruptions. Requests by industrial customers for voltage regulation to ±3/4 volt, and for deviations from normal voltage for not longer than one cycle have been reported. Such limits are not practical today. No limits of tolerance can be assigned to voltage dips caused by uncontrollable transients such as short circuits and faults. Limitation of voltage dips from controllable causes to from 2 to 6 volts (120-volt basis) is a matter of judgment depending on the frequency of occurrence and the class of service affected by the dips.

The permissible amount of flicker voltage cannot be stated concisely for a number of reasons. The voltage flicker is objectionable primarily because of the reaction of the human eye to variations in electric light intensity. The reaction of individuals varies over a wide range. Flicker objectionable to one may not even be perceptible to another. Lamps of different size and type produce unlike light flicker from the same voltage flicker. The magnitude of the voltage change, the rate of change, the duration of the change, and the frequency of occurrence of the flicker, all have an effect on the limit of flicker tolerance.

Numerous investigators have studied the flicker problem. The most complete analysis is found in a 1937 report by Utilities Coordinated Research, Inc.[1] Fig. 1 presents a chart generally accepted as the result of a compromise of allowable limits set up by various authorities for their own use. The most irritating light flicker is in the order of six dips a second.

Causes of Dips and Flicker

Short circuits and faults are unavoidable operational hazards on a central station system. They produce voltage dips on supply busses of substations proportional to the ratio of the current delivered to the short circuit from the supply bus and the current in a short circuit occurring on the supply bus itself. The duration of such dips is the combined relay and circuit-breaker time required to remove the faulted equipment from the system. Even the fastest and most modern relays and breakers have combined times in excess of three cycles. Many relatively new relays and breakers, which a utility cannot afford to retire, have combined times of 20 cycles or more for short circuits causing voltage dips of 30 per cent on the supply bus. From an economic standpoint alone it will be many years before the duration of voltage dips can be reduced materially.

Cases of voltage flicker originating on the power system itself are extremely rare and, when detected, are eliminated without delay. Most of the flicker on central station systems is due to the customer's utilization equipment. The most common voltage problems are caused by the starting of motors. For reasons of cost, efficiency, and reliability, motors are designed to require a momentary starting current several times their full load running current in order to produce sufficient starting torque. For approximate calculations, a starting current of six times normal at 35 per cent power factor may

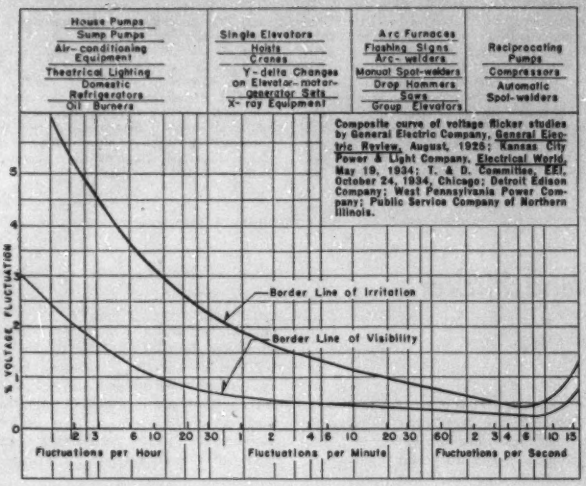


Fig. 1. Relations of voltage fluctuations to frequency of their occurrence

be used, but specific data should always be used when available. Motor-driven reciprocating loads such as compressors produce cyclic pulsation of voltage. Comparatively small variations of voltage may be objectionable if the pulsation occurs from 6 to 12 times per second.

Electric welders represent a class of equipment of great importance in power system flicker. Most welders have a smaller "on" period than "off" period, and consequently, have a low load factor. Fortunately, most welders are located in factories, where other processes require a large amount of power, and where the supply facilities are sufficiently heavy, so that no flicker trouble is experienced. In isolated cases the welder may be the major load in the area, and serious flicker may be imposed on the distribution system which otherwise may be adequate for ordinary loads.

Of the three types of electric furnaces, resistance, induction, and arc, the last frequently causes voltage flicker. The operation of an arc furnace has two periods, the melting down period and the refining period. During the melting down period, pieces of scrap will, from time to time, at random, bridge the electrodes and approximate a short circuit on the secondary side of the furnace transformer. Consequently, the melting down period is characterized by violent fluctuations of current at low power factor. After the steel has been melted down a stable arc can be maintained on all electrodes by automatic electrode regulators. The refining period is characterized by a steady load of high power factor. The utility's concern is, therefore, confined to the melting down period of the arc furnace operation and the voltage flicker produced during this period. This problem has received considerable study by utilities supplying service to arc furnaces. It has been recognized by the AIEE Committee on System Engineering and a working group on arc furnaces has been established to develop a guide for application of arc furnaces on power systems. A preliminary report was presented as a conference paper at the 1955 AIEE Fall General Meeting. The curves in Fig. 2 express the substance of the working group's preliminary recommendations. It is a zone chart indicating the effect of connecting an arc furnace of a given megavolt-ampere rating of its transformer to a supply bus of a known short-circuit capacity. No objectionable flicker will be produced on busses with short-circuit capacity higher than shown by both curves, while objectionable flicker is to be expected on busses with short-circuit capacity lower than the two curves. Short-circuit capacity lying between the two curves indicates marginal conditions where objectionable flicker may occur from time to time and complaints from other customers of the utility may or may not compel the application of corrective measures.

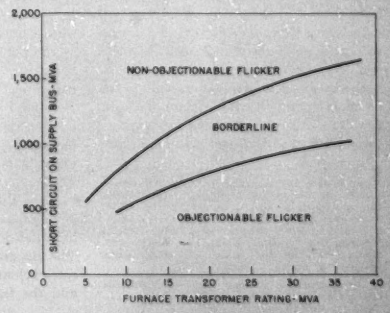


Fig. 2. Effect of electric arc furnaces on the voltage of the supply bus

The use of large continuous rolling mills for producing wide metal strip introduced new problems of speed and voltage regulation on utility systems. The loads, unless supplied through rectifiers, do not necessarily produce flicker in the sense discussed here. The load comes on and drops off in steps as the metal enters and leaves the rolls. The demand builds up in steps to full value in a few seconds, remains at full value for a longer period, then drops to almost zero in a few seconds. Whether or not a flicker takes place during this operation, there is nevertheless a voltage swing during the cycle of the rolling mill. The result is an excessive operation of voltage-regulating equipment. When it is large enough, a cycling load of this nature produces phase angle swings on the power system akin to the old problem of hunting.

Prevention and Cure

The foregoing briefly outlines the problems of voltage dips and flicker facing the utility engineers. As in matters of health and human relations, it is infinitely better to prevent the occurrence of trouble rather than to cure it after it has occurred. To do so it is essential that industrial en-

gineers designing the customer's plant and utility engineers planning the system extension to serve the customer have a complete understanding of each other's problem. The utility engineers must protect other customers supplied from their system against voltage dips and flicker in excess of the limits of tolerance yet they must do so at the lowest possible cost. They need to know the magnitude and duration of transient demands of the load to be supplied and methods of reducing them that may be available to the customer at a reasonable cost.

Since the magnitude of the voltage disturbance is measured by the ratio of the current causing the disturbance to the short-circuit current on the bus where the disturbance is observed, the obvious and fundamental solution would be to supply the customer from a bus having a sufficiently large short-circuit capacity to hold the ratio of currents within the limits of tolerance. The elements to be considered by the utility engineer are as follows: (1) detailed knowledge of the electrical characteristics of the customer's load; (2) limit of tolerance of voltage dip and flicker for this load; and (3) short-circuit capacity of substation busses within economic transmission distance from the customer's plant. Knowing the magnitude of current disturbances produced by the customer's equipment and the limit of tolerance of the corresponding voltage dips and flicker, the utility engineer can estimate the minimum short-circuit capacity of the required supply bus. He can then select the nearest substation bus having a short-circuit capacity in excess of the required minimum. Such a solution will result in satisfactory service to all concerned.

All too often, however, the solution is not as simple as outlined. Adequate information on operating characteristics of the customer's equipment is not always available or is grossly underestimated. Equipment that is potentially disturbing to the utility system is at times added in an existing customer plant which had caused no disturbances up to that time. There may not be a substation bus of adequate short-circuit capacity within reasonable transmission distance of the customer's plant. In such cases it becomes necessary for the utility engineers to call upon the industrial engineers for a joint study of remedial measures directed toward the reduction of the magnitude of disturbing currents and the reduction of their rate of change. In some cases industrial engineers may be able to minimize voltage dips and flicker by controlling manufacturing processes. Voltage dips caused by starting motors across the line may be reduced by the use of starters which limit the inrush of starting current and prolong the starting time. Intermittent 3-phase and single-phase loads may be supplied through motor-generator sets and flywheels may successfully and economically reduce disturbances caused by motor-driven reciprocating or intermittent loads.

Should the remedial measures under study fail to reduce disturbances to a satisfactory level to prove to be uneconomical, it would then be necessary for the utility engineers either to "stiffen" the system to the supply bus or run a separate transmission line to the customer from the remote bus of high short-circuit capacity. The difficulties in solving the problems of this kind are seldom technical in nature. It is usually the determination of the equitable division of the extra cost which is required to supply power to disturbing loads that requires prolonged negotiations.

Conclusions

In summing up the utility's problem of voltage dips and flicker with specific reference to industrial customers, the following conclusions are submitted:

1. Designers of industrial control systems and automatic machines should not assume continuous and constant voltage supply from the utility system. Utility engineers must tell the customers that while every effort is being made to minimize voltage dips, they cannot be eliminated.

2. When given complete information on the electrical characteristics of a prospective new load, and its ultimate size, planning engineers of the utility system will provide an adequate extension of their system to serve the customer and to prevent disturbances to other customers.

3. Ignorance or misinformation regarding transient or cyclic performance of customer equipment may result in voltage dips or flickers on the utility's distribution system disturbing to other customers and which would, therefore, require correction. At best, the determination of the equitable shares of cost of correcting disturbances is difficult.

Mutual understanding by industrial and utility engineers of each other's problems should go a long way toward improved service to all concerned.

References

1. The Visual Perception and Tolerance of Flicker. Utilities Coordinated Research, Inc., New York, N. Y., 1937.

2. Distribution Data Book. *GET-1001D* General Electric Company, Schenectady, N. Y. July 1951.

3. Large Electric Arc Furnaces—Performance and Power Supply, B. M. Jones, C. M. Stearns. *AIEE Transactions*, vol. 60, July 1941, pp. 763–69.

4. Arc-Furnace Loads on Long Transmission Lines, T. G. LeClair. *Ibid.*, vol. 59, Apr. 1940, pp. 234–42.

5. Voltage Dip and Automation, C. E. Quick. *Bulletin*, Edison Electric Institute, New York, N. Y., Oct. 1955.

Discussion

L. G. Levoy (General Electric Company, Schenectady, N. Y.): Mr. Kroneberg has nicely summarized the voltage dip and flicker problem. I would like to supplement his summary with a few additional comments.

We have found that some large continuous mills in steel plants may produce objectionable voltage disturbances on power systems even when supplied through synchronous motor-generator sets. This disturbance can be greatly reduced in magnitude by the use of power factor regulators on the motor-generator set motors. For those noncontinuous mills which require large power for short passes and where the duty cycle is low enough, wound-rotor induction motors with slip regulator and flywheel can be used effectively to minimize the disturbance to the power supply system.

In calculating the magnitude of the voltage disturbance on a system caused by a given change in load current, it is necessary to consider the power factor of the current causing the disturbance and the X/R ratio of the supply system as well as the short-circuit level at the bus. That is, the system voltage drop caused by the load current must be oriented vectorially with respect to the circuit voltage so that a realistic value is obtained.

In cases where a load is known to produce objectionable flicker voltage disturbances on the power supply system available, such as an arc furnace on a system which is not sufficiently "stiff," one alternate to "stiffening" the system or running a separate line to a bus of high capacity, is the use of a synchronous condenser and buffer reactor at the furnace bus to minimize the shock var loading on the supply system. In making an economic comparison of the various ways of successfully serving such a load, this method should also be considered.

C. E. Quick (The Detroit Edison Company, Detroit, Mich.): The author is to be commended for an excellent summation of a problem which is assuming increased importance to power company engineers and to their customers.

One note of caution should, however, be emphasized. The curves of Fig. 2 would indicate that a 25,000-kva furnace could be installed on a supply bus of 1,000,000-kva short circuit and be borderline between

objectionable and nonobjectionable flicker. This may be true for some 25,000-kva furnaces. However, tests[1] on two new 25,000-kva furnaces, with an electrode voltage of 510 volts, showed single-phase load swings of 11,000 kva four to six times per second and other single-phase load swings of 19,000 kva not oftener than once per second. Such swings would result in flicker on the 1,000,000-kva bus of more than 2 volts at the cyclic rate considered most irritating to the eye, and of nearly 4 volts at an extremely frequent rate. Both of these flickers are highly objectionable when referred to the curves of Fig. 1. Since the trend in electric furnace installations is toward both larger sizes and higher electrode voltages, considerable caution should be exercised in engineering for new furnaces. The use of data from smaller furnaces with lower electrode voltages will certainly result in objectionable flicker.

REFERENCE

1. TWO LARGE ELECTRIC ARC FURNACES—ELECTRICAL CHARACTERISTICS AND CORRECTIVE EQUIPMENT, S. W. Luther, J. D. Ghesquiere, C. E. Quick. *AIEE Transactions*, vol. 74, pt. III, 1955 (Feb. 1956 section), pp. 1401–07.

Robert F. Lawrence (Westinghouse Electric Corporation, East Pittsburgh, Pa.): As original chairman of the AIEE Working Group on Arc Furnaces, I believe it is wise to emphasize the preliminary status of the group's work as reported in the conference paper mentioned in Mr. Kroneberg's paper, and to which an added note of caution is in order. The results thus far published do not constitute strict definitive limits. Although the present report is under revision, it is still unlikely because of the nature of the problem that a "black and white" zone can be established.

The author has outlined the effect of furnace operation on system voltage. Some of the working group members have raised the question of the effect of cyclic flicker, i.e., voltage changes of a frequency of five to six cycles per second. Cyclic flicker can constitute a criterion of objective flicker, especially on large furnaces, 25,000 kva and larger, and therefore should be investigated. In the arc furnace survey the general criterion of "objectionable flicker" was used and defined as: "Means customer complaints. Describe in a separate memorandum any qualifying statements which you feel should accompany the answer to this column." Qualifying statements were included in the preliminary report.

The curve of Fig. 2 of the paper expresses the substance of Fig. 11 of the conference paper with the exceptions noted in the interpretation that follows. In the arc furnace paper, the term mutual drop constant was expressed as a function of the arc furnace transformer rating in megavolt-amperes. Mutual drop constant is merely a measure of the voltage drop from the source to a point in a grid, due to a current flow from another point in the grid. Thus, rated arc furnace transformer current flow will produce a per-unit voltage drop to a point in the electric supply system of the magnitude of the mutual drop constant, Xmn (expressed in per unit on the furnace transformer rating in megavolt-amperes).

In the conference paper, the term self-drop constant was also used. Self-drop constant is a measure of the voltage drop from the source to the point in the grid due to current flow from that point. Thus, self-drop constant (Xmn) expressed in proper per-unit terms is merely short-circuit impedance or it is a special case of the mutual drop constant. For the case of an arc furnace transformer and a lighting load located at the same point, the self- and mutual drop constants are equal. The location of the furnace with respect to the location of the lighting load is not specified in the report. Almost exclusively they are not the same.

The ordinate of Fig. 2 of the paper is actually obtained by dividing transformer rating of the curve of Fig. 11 by the corresponding value of Xmn in per unit.

The author has chosen to call this ordinate "short circuit on supply bus—megavolt-amperes." This nomenclature is correct only if the lighting load and the arc furnace load are at the same location. The ordinate of Fig. 2, previously stated as the reciprocal of the mutual drop constant Xmn, is a measure in terms of megavolt-amperes of the stiffness of the system to the point of the lighting load regardless of the location of the arc furnace load. Thus, a better name for the ordinate might be "stiffness index—megavolt-amperes."

A. A. Kroneberg. The discussions supplement the paper and are much appreciated. I would like, however, to defend the use of the term "short circuit" for the ordinate of Fig. 2. The short-circuit capacity of a bus in itself is an index of stiffness. It is a quantity that must be known to a planner, a designer, and an operator. Circuit breakers connected to the bus must have rupturing capacity in excess of the short-circuit capacity of the bus. Protective relays are installed and set to operate on the basis of short-circuit capacities of substation busses. When static capacitors are installed to correct low-voltage conditions on a substation bus, the resulting voltage boost is measured by the ratio of the capacitive kilovolt-amperes to the short-circuit capacity of the bus. These and many other reasons explain why the term "short-circuit capacity" is well known among electrical engineers and why there is a desire to express results of system studies in terms of short-circuit capacity as in Fig. E-1 of the preliminary report, in Fig. 2 of this paper, and in both illustrations of reference 1.

The discussions cautioning on the use of Fig. 2 of the paper tend to emphasize my statements elsewhere in the paper that "adequate information on operating characteristics of customer's equipment is not always available or is grossly underestimated" and that "ignorance or misinformation regarding transient or cyclic performance of customer equipment may result in voltage dips or flicker on the utility's distribution system."

REFERENCE

1. SYSTEM CAPACITY REQUIRED FOR 3-PHASE ARC FURNACES, Oliver Ramsaur, J. E. Trowbridge. *AIEE Transactions*, vol. 73, pt. III-B, Dec. 1954, pp. 1373–74.

Rapid Transit Progress in Toronto

J. G. INGLIS
MEMBER AIEE

CANADA'S first subway began operating in Toronto on March 30, 1954. In the first 2 years of operation it has carried about 144,000,000 passengers and its cars have travelled about 12,000,000 miles. So successful has it been that the initial order of 106 cars is being increased over 30% (per cent) by an order for 34 more. Their arrival during the coming summer will permit the system to operate at its ultimate designed capacity of 40,000 passengers per hour in each direction. Its popularity with the riding public has created an urgent demand for a second subway and has proved the advantages of rapid transit in one more metropolitan area.

Paper 56-762, recommended by the AIEE Land Transportation Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Summer and Pacific General Meeting, San Francisco, Calif., June 25–29, 1956. Manuscript submitted March 27, 1956; made available for printing August 20, 1956.

J. G. INGLIS is with the Toronto Transit Commission, Toronto, Ont., Canada.

General

The subway of the Toronto Transit Commission operates under or near the city's main thoroughfare, Yonge Street, connecting the downtown, hotel, railway, and lake-front area with the northerly part of the city. The line is 4.5 miles in length, of which about two-thirds, including all of the downtown area, is under ground, while the remaining one-third is above ground.

The subway is the backbone of the city's transportation system, making free transfer connections with 27 streetcar, trolley coach, and bus routes. With its fleet soon to be increased to 140 subway cars, 5,500 Commission employees provide transportation service for the 1,250,-

000 inhabitants who live in the 240-square-mile-area of metropolitan Toronto. Other equipment includes 885 streetcars, 437 busses, 140 trolley coaches, 5 steam and Diesel ferry boats, and an interurban coach subsidiary covering 1,450 miles of highway with 222 vehicles.

There can be little doubt regarding the need for the subway in Toronto. Yonge Street previously had the distinction of having the most heavily travelled streetcar line in the world. It also had the distinction of being the last route on this continent that had streetcars and trailers which, during and after the war, carried as many as 12,500 passengers per hour in the rush periods. The same cars and trailers, roughly 30 years of age, were operated on worn-out track. At the end of 90 years of streetcar service, the schedule speed in the downtown area could barely exceed the 6-mph (miles per hour) limit imposed by its original charter. Add to this a roadbed on a narrow street having a 1¼% grade against outbound traffic with its attendant delays and tie-ups during slippery conditions, and there can be no doubt that a drastic improvement was needed.

The Car

The first order of 106 cars was placed in England following competitive bidding from manufacturers in the United States, Canada, Great Britain, France, and Belgium. The 34 cars currently on order are being built by the same firm and are practically duplicates of the original ones. Their purchase in Great Britain helps to pay for the export of Canadian products from farm, forest, and mine.

The dimensions of the cars are:

Car length, over bumpers, 57 feet 1½ inches
Car length, over body, 55 feet 7½ inches
Car width, 10 feet 4 inches
Car height, 12 feet 0 inches
Truck centers, 38 feet, 0 inches
Truck wheelbase, 7 feet 0 inches
Wheel diameter, 30 inches
Door openings per car side, 3
Door opening width, 3 feet 9 inches
Door height, 6 feet 4 inches
Seating capacity, 62

The cars are semipermanently connected in pairs to form 2-car trains to reduce first cost, weight, and operating expense. Since the minimum train size has been found to be four cars, the cars on order will have no driving positions but will also be connected in pairs in the middle of existing 2-car units to form 4-car trains, thereby providing further reduction in weight and expense.

The trains vary in size between four, six, and eight cars in accordance with the traffic demands. In the rush hours, 8-car trains can operate on a 2-minute headway to carry 40,000 persons per hour with 167 passengers per car. At the present time, the rush-hour headway is 2½ minutes and the train size is largely confined to 6 cars, pending arrival of the cars now on order. The chief purpose of this paper is to focus attention upon the various effects of the first rapid transit system built in metropolitan Toronto.

Speed and Traffic

The subway has produced a revolution in travel time and convenience over the area served, not unlike that produced by the telephone or the electric light.

From the terminal of the former streetcar line, 1.9 miles beyond the subway terminal, the streetcar schedule was 40 minutes; increases of 25% or more were not uncommon. Today the comparable time by bus and subway has been cut to 26 minutes, and the timetable can be closely adhered to in good weather and in bad. During the evening rush, on the day of Hurricane Hazel, the subway performance was almost normal.

Richmond Hill, a town 10½ miles beyond the city limits, used to be 80 minutes from downtown Toronto by bus and streetcar. Today, because the subway operates a partial express service, the running time is only 64 minutes, a decrease of 20%, while the number of passengers has increased by over 25%.

The increased capacity of the streets for motor traffic, and of the subway for transit riders, is shown in Table I. This table compares the conditions existing in February 1954, on downtown Yonge Street, just before streetcars stopped operating, with subway operation 2 years later.

Table I. Comparison of Conditions

Daily Passengers Entering Downtown	February 1954	February 1956	Ratio 1956/1954
Public transit	67,720	151,337	2.23
Automobiles	11,755	20,664	1.76
Combined	79,475	172,001	2.17

The convenience of the subway is a decided factor in its popularity. Even though the cars are jammed in rush hours, and a seat is the exception rather than the rule, the quiet ride without the annoyances and delays of street traffic is a great attraction for persons of all occupations. A survey conducted in March 1956 revealed that slightly over 10% of the passengers boarding the subway at its outer terminal formerly travelled by automobile.

The enthusiastic reception of the subway by its users means just one thing—give the public what it wants and it will react accordingly, whether it be to "hi-fi," stretchy nylon sox, the chance to win $64,000 on a television program, or good service in getting to and from work. Those who provide transportation have an obligation to provide service that is safe, comfortable, economical, and fast. They must give due attention to all four, but if they should ask their passengers which one they consider to be most important undoubtedly the vast majority would say speed. The Toronto experience has proved in yet another city that fast, dependable, economical transit service is a positive rebuttal to those who believe that bigger and better facilities for the automobile are the only means of moving workers in a metropolitan area.

Selling Transportation

Transportation companies are in business to sell transportation or rides. The number of rides sold will be in proportion to the value of the commodity offered for sale, exactly as in the case of any other article which is purchased, and a subway service is no exception.

Furthermore, there is no clear line of division whereby certain features of the sales appeal of the product are engineering matters while others are not. Instead, all are very important to the engineer and it is partly the purpose of this paper to stimulate Institute members to apply their knowledge and viewpoint to all problems of a rapid transit system far beyond the technical ones.

There are two fundamental factors in plain evidence in the Toronto subway, both important and related to each other. They are, first, the sales appeal of the ride or service and, second, the factors which make the service socially acceptable to all.

Many things have been done to obtain the maximum schedule speed. When the line first started, in March 1954, the round-trip running time, including layover, was 40 minutes. After a little experience it was reduced to 37.5, then to 36. It was reduced further to 34 by a "drop-back" system used in rush hours only, by which, when a train reaches the end of the line, the motorman from the previous train is standing at the point where the last car stops, ready to operate it in the reverse direction; this is possible in 30 seconds when necessary. The reduction in layover time possible by the "drop-back" system permits the base

round-trip time of 36 minutes to be reduced to 34 in the rush hours, 15.88 mph with stations spaced 0.4 mile apart.

Another factor contributing to increased schedule speed is the use of a whistle by the guards; two blasts indicate the closing of the car doors at stations. There are two desirable results, namely, increased safety of passengers who avoid the doorways as soon as the whistle is blown, and a slight reduction in the stopping time.

Trains are dispatched automatically from each terminal by means of a tape driven at a constant speed over a light-sensitive relay. This eliminates any human error in dispatching them and promptly helps to avoid delays.

Many things, in addition to speed, have been done to make the ride attractive. Color has played an important part in this respect. The station walls are covered with Vitralite or glass-faced masonry, colors of yellow, turquoise, and gray being used in successive stations with contrasting trim colors and with matching color in the terazzo floors. The different wall colors also serve to identify further the stations. Frequent cleaning and absence of horizontal dust-collecting ledges further enhance the appearance. The public response to this use of color has been enthusiastic. A somewhat similar result was achieved with the cars. Their exterior is the Toronto Transit Commission's standard bright red with gold trim, and the trains are put through an automatic car washer often enough to keep them clean. The car interior has bright and complementary colors in the ceiling, walls, seats, and floors. These colors, as well as plastic surfaces, liberal use of bright metal finish, frequent cleaning, and the deliberate avoidance of dust pockets increase the sales appeal.

Great emphasis has been placed upon quietness of operation, as indeed it must be when the chief competitor is the motorcar with its many desirable features including quietness. Acoustic ceilings have been used in stations wherever feasible, chiefly in the mezzanine areas and stairways, but in train platforms as well.

Noise reduction in the tunnel also received special attention. The common, or centre, wall in the tunnel is covered, on both sides, by a 2-inch thickness of mineral wool 4 feet high. The same thickness of insulation is also applied on the vertical wall under the station platform. This insulation is believed to be the most extensive in any subway in the world and has achieved the objective of reducing noise to tolerable levels and of permitting conversations to be carried on without difficulty. Its continued use is expected in future subways.

Noise reduction in the cars also received special attention. The drive is through hypoid gearing and is very quiet. The transmission of external noises into the car is reduced by a 1/4-inch rubber floor on top of $1^1/_2$ inches of cork. There is also an attachment of a heavy cloth fabric, on the inside of the body shell, from the floor to the window-sill level. Coupler noise has been reduced by a spring-loaded buffer plate which keeps the couplers in tension, thereby minimizing lost motion.

Passenger convenience was given a high priority. The illumination of the cars is approximately 20 foot-candles, which permits reading in comfort. The car temperature is under thermostatic control with forced-air heating at sufficiently low velocity as not to be felt. The seating arrangement is a combination of longitudinal and cross seats so arranged that no seated passenger is much inconvenienced by the movement of another. Three double doors on each car side put every passenger within five steps of the nearest door. Stanchions and ceiling-mounted standee grips in adequate supply, have combined utility with beauty. The subway has been placed as close to the surface as possible to reduce stairway heights, while 15 escalators provide aid at the points of greatest heights or the heaviest traffic. Transfers between subway and surface vehicles have been made as simple and convenient as possible. Wherever possible, the transfer to or from the surface vehicle is made on Toronto Transit Commission property thereby eliminating any need for a paper transfer. This permits the loading and unloading to be done via all doors and in minimum time. A public address system at all stations informs the passengers quickly of any item of importance, such as a delay, and the same system is used to transmit general or coded messages to the employees.

Buildings and Assessment

The subway has pioneered a building boom along the route. New office buildings are being erected on most of the vacant land or as replacements for former old buildings razed to provide space for the new. Apartment houses are being erected either close to the route or within walking distance of it. A group of insurance firms has erected or enlarged its offices on one of the main cross streets, also within walking distance. Parking lots have been built over some of the open cut sections and buildings are being considered also. And some areas which could properly be described as blighted are being renovated. The exact effect of all these changes on assessment and taxes has never been accurately determined and is indeed very difficult, but in the immediate areas values have doubled or tripled, while a visual inspection of the new construction is most convincing evidence of the new life blood which speedy mass transportation can pump into a metropolitan area.

Eglinton Terminal

The subway ends at Eglinton Terminal which is not unusual in construction or function, but it provides a lesson, much needed, at times, in the benefits of mass transportation and traffic control on city thoroughfares. The terminal, at the southwest corner of Eglinton Avenue and Yonge Street, has nine bus and trolley coach routes operating in three directions to and from it. About 200 vehicles per hour enter the terminal during rush periods, or one entering or leaving every 10 seconds. There are no left turns at the intersection and no stopping allowed in the immediate area, all of which has combined to produce a service that has proved those pessimists to be wrong who said that chaos would result if so many busses were operated to and from one spot. It has also given a convincing demonstration of the capacity of streets when people travel by public transit rather than at the rate of 1.7 people per automobile.

Conclusions

Canada's first subway has justified the wisdom and courage of those who conceived, planned, and built the system. It has caused a revolution in travel for about 125,000 persons by providing them with a speedy, economical, comfortable, and safe transportation means which is relatively immune to traffic and weather. It has been accepted enthusiastically. It has increased assessment and taxes and will continue to do so for years to come. It has caused municipal authorities to reconsider the order of priority among future expressways and subway lines. It has caused an increasing demand for a second, or east-west, subway to be built at the earliest possible date. The Chairman of the Metropolitan Council stated in 1956: "We know that beyond a certain stage 1 dollar spent on rapid transit is worth 5 dollars spent on more arterial highways and parking facilities."

Analysis of Tandem Cold Reduction Mill with Automatic Gauge Control

R. A. PHILLIPS
MEMBER AIEE

MOST of the cold rolled steel strip which is produced today is rolled on tandem cold reduction mills. Fig. 1 shows a 5-stand mill with its unwind and winding reels. The strip fed into the mill has previously been hot rolled and pickled. Each stand, using a combination of pressure transmitted through the work rolls and interstand tension, reduces the strip thickness by as much as 50 per cent. When the delivered strip is later to be plated and made into tin cans, the nominal input thickness to this mill may be 90 mils and the delivered thickness 8 mils. Typical delivery speeds for this product are 3,000 feet per minute up to nearly 6,000 feet per minute. The subsequent operations upon the steel to obtain the finished product require that each coil of strip be rolled to close thickness tolerance. Although mill operators can and do control thickness by manual adjustment, this is difficult and tedious even for the most alert operators. Therefore, a completely integrated automatic gauge (thickness) control for a tandem cold strip mill has been built and placed in service for more than a year. Other equipments are being manufactured and some are being installed on other mills.

This paper describes the control system, develops the transient equations for describing mathematically the tandem mill rolling process, and develops the block diagram from which a computer study is made to study gauge control performance.

Description of Integrated Gauge Control System

The first installation was made on an existing 5-stand tandem cold mill. Over a period of years the mill had been the subject of considerable testing which made adequate data for system studies available. The rated delivery speed of the mill is 3,850 feet per minute. The individual stand drive motors are fed from a common bus. The automatic gauge control system has been described previously in considerable detail[1] so will be described here only briefly.

A "coarse" regulating system was installed to operate on the screwdown motors of stand *1*. Fig. 2 is a functional diagram of the system. The strip, as delivered from the hot strip mill will have rather large thickness variations in it. Even though the mill housing, the work rolls, and the backup rolls are "stiff," the opening between the work rolls will vary since the variations in incoming gauge will change the forces required. Although it may seem that such variations must be small, it is precisely this effect which makes gauge control necessary. An X-ray thickness gauge is located at the delivery side of stand *1*. When the desired strip thickness has been set, as a reference, on the X-ray gauge, an electrical error signal is developed proportional to deviation from the reference. This signal amplified by transistors and magnetic amplifiers, operates the screwdown motor control for an "on" time proportional to gauge deviation error.

Due to the physical construction of mills of this type, it is not possible to locate the X-ray gauge closer to the work-roll pass line than 6 feet. Time then must elapse between when the material is rolled and when the X-ray gauge measures it. This transport time is a minimum of 1/2 second and, when a coil is being threaded into the mill, may be 5 seconds. Once a correction is made, the control must wait for this transport time to elapse before it can decide whether another correction is required. A sampling-type system is therefore a logical selection. The "off" time must be greater than this transport time plus the sum of any other delays in the circuit. The transport time is inversely proportional to strip speed. This calculation is the extent of the analytical work required to engineer a coarse control.

The stand *1* coarse control then has the function of ironing out variations produced by the hot mill. It is important that the stand *1* control does not overcorrect and thereby create thickness errors. The subject system has been adjusted so there is no overcorrection. Since there are four more stands through which the strip must pass before it leaves the mill, it can be expected that vernier correction will be desirable in stand *5*.

It is a well-known fact to tandem cold strip mill operators that, with a winding reel control acting to hold constant winding tension, a change in stand *5* speed will produce a change in delivered gauge. If the speed of stand *5* is increased, the tension will increase between stands *4* and *5*. This increase in tension will decrease the delivered gauge. (The theoretical approach given later in the paper will treat this subject much more thoroughly and exactly.)

Since each of the stands is reducing the strip in thickness, the outgoing speed of a stand is much greater than the incoming speed. As a consequence, the transport time from the stand *5* work-roll pass line to the X-ray gauge 6 feet away ranges from about 0.1 second to a maximum of 0.5 second. A continuous-type vernier control is then practical to use.

With the common bus arrangement, the speed of the motors may be varied by controlling field current. The stand *5* motor is provided with a fast-acting amplidyne-excited field-current regulator. During automatic operation the error signal of the stand *5* X-ray gauge is amplified by transistors and magnetic amplifiers to increase or decrease the excitation of the amplidyne. This then verniers the speed of the stand *5* drive motor. Fig. 3 is a functional diagram of the vernier system.

It will be shown that it is necessary to describe mathematically the physics of the rolling process itself since the stands and winding reel are coupled together by tension and by propagation of strip thickness changes. Therefore, an analysis of the vernier control must consider not only the transient characteristics of the stand *5* motor and control, but also must consider the behavior of the other stands and the winding reel as well. The transient performance characteristics of the winding reel are most influential since the reel is on the delivery side of the continuously controlled stand, stand *5*.

Derivation of 5-Stand Tandem Cold Mill Equations

From experience in rolling mills it is known that the thickness delivered from a stand can be affected greatly by changes in rolling pressures due to operating the screwdowns, by changes in motor speed, and by variations in input thickness. In addition, torque requirement from a

Paper 56-638, recommended by the AIEE Mining and Metal Industry Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Summer and Pacific General Meeting, San Francisco, Calif., June 25–29, 1956. Manuscript submitted March 29, 1956; made available for printing April 17, 1956.

R. A. PHILLIPS is with the General Electric Company, Schenectady, N. Y.

Fig. 1. 5-stand tandem cold rolled strip mill with automatic gauge control

stand is affected by these changes. The torque requirement, in turn, influences motor speed. It is also known that "stiffness" or "spring" of the stands of one mill may be different from that of the stands of another mill, resulting in somewhat different rolling characteristics.

An analysis then should be in two separate parts in the beginning: (1) the equations which apply to the physics of rolling only; and (2) the equations which describe the pertinent characteristics of a particular mill. They should be kept separate as long as is practical. This makes it possible to study the effect of rolling the same product on different mills.

In all the derivations presented in this paper it is assumed that there is tension between stands and that the material is elastic between stands. In a tandem cold reduction mill tension is maintained at all times so all cases of interest are described properly by this assumption.

In Appendix I the equations relating interstand tension to delivery speed of one stand and to entry speed of the next are derived.

$$\frac{T_1}{A_1}=\frac{E}{L_4}\int_0^t [v_{21}-v_{10}]\,dt \tag{1}$$

$$\frac{T_2}{A_2}=\frac{E}{L_4}\int_0^t [v_{32}-v_{21}]\,dt \tag{2}$$

$$\frac{T_3}{A_3}=\frac{E}{L_4}\int_0^t [v_{43}-v_{32}]\,dt \tag{3}$$

$$\frac{T_4}{A_4}=\frac{E}{L_4}\int_0^t [v_{54}-v_{43}]\,dt \tag{4}$$

$$\frac{T_5}{A_5}=\frac{E}{L_R}\int_0^t [v_R-v_{54}]\,dt \tag{5}$$

No Deformation in the Roll Gap

To understand the effects of plastic deformation of the strip between the work rolls, it is desirable first to study the case where no plastic reduction is being taken. A distinction should be made here between elastic and plastic deformation. Elastic deformation is recovered upon relief of stress; plastic deformation is permanent in nature. If all deformation were elastic it would be possible to pass strip through a tandem cold mill with substantial interstand tensions without permanently changing its thickness. While the strip is in the mill it would be temporarily thinner due to elastic deformation. In Appendix II the velocity of the strip delivered from a stand is derived in terms of incoming and outgoing tensions for this case. As is pointed out in Appendix II, incoming velocity to a stand is the roll peripheral velocity so v_{21}, v_{32}, v_{43}, v_{54}, of equations 1–4 become v_{20}, v_{30}, v_{40}, and v_{50}, respectively. If the equations are expressed in operational form, later manipulations will be made more simply.

Equation 65 of Appendix II may be substituted into the thus-modified equation 4 and similar equations substituted into equations 1–3 and 5.

$$p\frac{T_1}{A_1}=\frac{E}{L_4}\left[v_{20}-\left(1+\frac{T_1}{A_1E}\right)v_{10}\right] \tag{6}$$

$$p\frac{T_2}{A_2}=\frac{E}{L_4}\left[v_{30}-\left(1+\frac{T_2}{A_2E}-\frac{T_1}{A_1E}\right)v_{20}\right] \tag{7}$$

$$p\frac{T_3}{A_3}=\frac{E}{L_4}\left[v_{40}-\left(1+\frac{T_3}{A_3E}-\frac{T_2}{A_2E}\right)v_{30}\right] \tag{8}$$

$$p\frac{T_4}{A_4}=\frac{E}{L_4}\left[v_{50}-\left(1+\frac{T_4}{A_4E}-\frac{T_3}{A_3E}\right)v_{40}\right] \tag{9}$$

$$p\frac{T_5}{A_5}=\frac{E}{L_R}\left[v_R-\left(1+\frac{T_5}{A_5E}-\frac{T_4}{A_4E}\right)v_{50}\right] \tag{10}$$

Although equations 6–10 do relate tensions to stand speeds in a tandem cold mill that is taking no reduction, it is difficult to study the operating characteristics with them in this form. Much can be learned if the analysis is limited to small excursions in speed about a steady-state speed.

If a reference point is taken such that interstand tensions are all zero and there is no slack between stands, all of the stands speeds will be equal. Call this speed v_0. Then

$$\left.\begin{aligned} v_{10}&=v_0+\Delta v_{10}\\ v_{20}&=v_0+\Delta v_{20}\\ v_{30}&=v_0+\Delta v_{30}\\ v_{40}&=v_0+\Delta v_{40}\\ v_{50}&=v_0+\Delta v_{50}\end{aligned}\right\} \tag{11}$$

If equations 11 are substituted into equations 6–10 and the results solved for tensions, the following are obtained:

$$\frac{T_1}{A_1}=\frac{E/v_0}{1+p(L_4/v_0)}[\Delta v_{20}-\Delta v_{10}] \tag{12}$$

$$\frac{T_2}{A_2}=\frac{E/v_0}{1+p(L_4/v_0)}\left[\Delta v_{30}-\Delta v_{20}+\frac{v_0}{E}\frac{T_1}{A_1}\right] \tag{13}$$

$$\frac{T_3}{A_3}=\frac{E/v_0}{1+p(L_4/v_0)}\left[\Delta v_{40}-\Delta v_{30}+\frac{v_0}{E}\frac{T_2}{A_2}\right] \tag{14}$$

$$\frac{T_4}{A_4}=\frac{E/v_0}{1+p(L_4/v_0)}\left[\Delta v_{50}-\Delta v_{40}+\frac{v_0}{E}\frac{T_3}{A_3}\right] \tag{15}$$

$$\frac{T_5}{A_5}=\frac{E/v_0}{1+p(L/v_0)}\left[\Delta v_R-\Delta v_{50}+\frac{v_0}{E}\frac{T_4}{A_4}\right] \tag{16}$$

From equation 12 it can be readily seen that a change in stand *1* speed will produce, through a time constant equal to the propagation time from stand *1* to stand *2*, a tension between stand *1* and stand *2*. It can be seen from equation 13 that this change in tension between stand *1* and stand *2* will produce an equal change in steady-state tension between stands *2* and *3*. Equations 14, 15, and 16 show that this same tension change will eventually be felt in all stands of the mill. From equation 16 it can be seen that a change in reel speed will change the steady-state tension between stand *5* and the reel only. One might say that in such a mill a tension change "works its way down the mill" but cannot work its way back. This is well known in the printing and paper industries where mills do not take reduction. However, when reduction is taken in each stand, the analysis becomes much more complex. Equations 11–16 will become useful for comparison.

Plastic Deformation in the Roll Gap

It is not the purpose of this paper to go thoroughly into the theory of plastic flow, but some description is necessary. It is assumed that throughout the roll gap the material flows plastically. Strip is made to flow plastically by squeezing it between the work rolls. If, in addition, tension is applied to the strip as it enters or leaves the roll gap, the roll pressure required will lessen since plastic flow theory says that the compressive stress plus the tensile stress must equal the yield stress at every point in the roll gap.[3] The rolls and the mill housing act like a large spring and tests show the combination obeys Hooke's law.[3] Therefore, when ten-

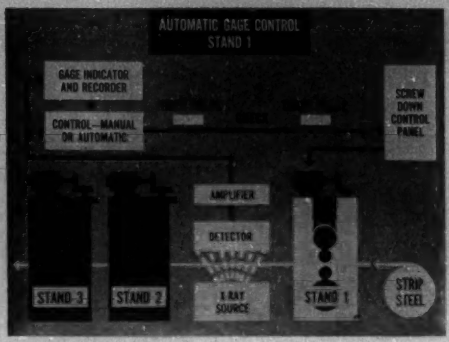


Fig. 2. Functional diagram of "coarse" gauge regulating system, stand 1

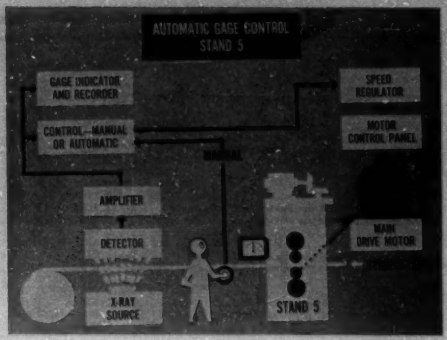


Fig. 3. Functional diagram of continuous vernier gauge control, stand 5

sion is applied, the roll pressure lessens and the work rolls move together until a new point of equilibrium is reached where compressive stress plus tensile stress equals the yield stress. The delivered thickness of the strip thus changes. This suggests that outgoing thickness can be changed by varying tensions or by changing roll pressure. The former is accomplished by speed changes; the latter is accomplished by moving jack screws fastened to the mill housing and pressing against the backup roll bearings. Rolling experience confirms this.

If it is assumed that cross sections of strip remain plane when passing through the roll gap, it follows that when reduction is being taken, there is one point on the roll arc of contact where the roll peripheral velocity equals the strip velocity. This will be called the neutral point. On the entry side of this point the rolls are moving faster than the strip; on the delivery side of this point the strip is moving faster than the rolls. The latter effect is called extrusion. This neutral point is, in general, between the entry and exit of the roll gap. An increase of incoming tension causes the point to move toward the exit; an increase of outgoing tension causes the point to move toward the entry. That these observations are correct can be proved by a physical picture.

Suppose the incoming tension were increased gradually. The work rolls would encounter an increasingly difficult job of pulling the strip through and would eventually begin to slip. Under this condition the rolls would be moving faster than the strip throughout the entire roll gap. This is, of course, the limiting case of the phenomenon described in the foregoing.

Changes in roll pressures also have effect upon the position of the neutral point. The torque (or power) required to reduce the strip is also influenced by changes in tensions and pressure.

The equations which relate the variables are nonlinear and extremely complicated. Fortunately, for the small changes which occur, the coefficients of linearized equations may be obtained numerically using the nonlinear equations. The principal causes of changes in roll separating force F, rolling torque, and neutral point thickness are changes in incoming and outgoing tensions and changes in incoming and outgoing thickness.

Assume that, for a particular stand, equations of this form can be written.

$$\frac{\Delta F}{F} = K_1\frac{\Delta h_o}{h_o} + K_2\frac{\Delta T_i}{A_i} + K_3\frac{\Delta T_o}{A_o} + K_4\frac{\Delta h_i}{h_i} \quad (17)$$

$$\frac{\Delta r_r}{r_r} = K_5\frac{\Delta h_o}{h_o} + K_6\frac{\Delta T_i}{A_i} + K_7\frac{\Delta T_o}{A_o} + K_8\frac{\Delta h_i}{h_i} \quad (18)$$

$$\frac{\Delta h_n}{h_n} = K_9\frac{\Delta h_o}{h_o} + K_{10}\frac{\Delta T_i}{A_i} + K_{11}\frac{\Delta T_o}{A_o} + K_{12}\frac{\Delta h_i}{h_i} \quad (19)$$

where

F = total roll separating force, millions of pounds
r_r = rolling torque, pound-feet
h_i = nominal incoming strip thickness, inches (measured under initial tension)
h_o = nominal outgoing strip thickness, inches (measured under initial tension)
h_n = neutral point thickness with initial tensions applied, inches
T_i = incoming tension, millions of pounds
T_o = outgoing tension, millions of pounds
A_i = incoming cross-sectional area, square inches
A_o = outgoing cross-sectional area, square inches
Δ = changes from the nominal or initial values

The nonlinear equations from which the K's were calculated were those derived by Bland and Ford.[4] The K's were found in the following manner: force F, torque r_r and neutral point thickness h_n were calculated for an initial condition. Then one of the conditions, say h_o was changed a few per cent. A new force, torque, and neutral point thickness were calculated and from them K_1, K_5, and K_9 were determined. A large-scale digital computer was used for these calculations.

Table I shows a typical tin-plate rolling schedule and the values calculated from it are shown to three decimal places in most cases. It is not claimed that the calculation method itself is this accurate, but the digital computer does all of its calculation with eight significant figures. Since this table gives intermediate results, it would not be wise to round off the numbers too far. The safest thing is always to carry more significant figures than required. The first five quantities are measured values; the remaining are calculated. Since the unwind reel does not supply tension, the constants K_3, K_6, and K_{10} have no significance for the first stand.

How much the roll separating force F changes is dependent upon the combined spring of the mill housing, screws, backup rolls, and work rolls, and the change in outgoing gauge and the screw setting change if any. In equation form

$$\Delta F = M[\Delta h_o - \Delta S] \quad (20)$$

where

M = combined spring constant, millions of pounds per inch
ΔS = change in screw setting, positive upward

Normally an equation describing a system of the type described by equation 20

Table I. Typical Tin-Plate Rolling Schedule

	Stand				
	1	2	3	4	5
h_t, inches	0.0900	0.0630	0.0390	0.0230	0.0140
h_o, inches	0.0630	0.0390	0.0230	0.0140	0.0080
T_t/A_t, millions of pounds per square inch	0	0.010	0.015	0.015	0.015
T_o/A_t, millions of pounds per square inch	0.010	0.015	0.015	0.015	0.009
F, millions of pounds	2.328	2.623	2.735	2.782	2.644
h_m, inches	0.0671	0.0410	0.0246	0.0150	0.00859
v_r, pound-feet	80,200	84,600	65,700	43,700	30,000
K_1	−2.488	−1.336	−1.204	−1.368	−1.275
K_2, per million pounds per square inch		−11.219	−11.390	−12.419	−13.729
K_3, per million pounds per square inch	−7.196	−5.394	−5.361	−6.692	−6.674
K_4	2.484	0.782	0.643	0.477	−0.002
K_5	−3.507	−2.007	−1.799	−2.019	−1.801
K_6, per million pounds per square inch		−12.206	−11.364	−11.006	−11.306
K_7, per million pounds per square inch	−3.358	−2.090	−2.081	−2.272	−2.741
K_8	4.205	2.152	2.066	2.170	1.744
K_9	0.889	0.916	0.906	0.892	0.909
K_{10}, per million pounds per square inch		−1.308	−1.121	−0.956	−1.052
K_{11}, per million pounds per square inch	1.171	1.105	0.882	0.716	0.742
K_{12}	0.0970	0.0460	0.0475	0.0615	0.0277

is a second-order differential equation. The inertia and damping terms are neglected because the natural frequency of the mill housing is very high compared to the natural frequency of any one of the regulating systems applied to a tandem mill.

If equation 20 is substituted into equation 17 and the result solved for $\Delta h_o/h_o$, the result may be substituted into equations 18 and 19 to give the following set of equations:

$$\frac{\Delta h_o}{h_o}=\frac{K_2}{\frac{Mh_o}{F}-K_1}\frac{\Delta T_t}{A_t}+\frac{K_3}{\frac{Mh_o}{F}-K_1}\frac{\Delta T_o}{A_o}+\frac{K_4}{\frac{Mh_o}{F}-K_1}\frac{\Delta h_t}{h_t}+\frac{M/F}{\frac{Mh_o}{F}-K_1}\Delta S \quad (21)$$

$$\frac{\Delta\tau_r}{\tau_r}=\left[\frac{K_2K_8}{\frac{Mh_o}{F}-K_1}+K_6\right]\frac{\Delta T_t}{A_t}+\left[\frac{K_3K_8}{\frac{Mh_o}{F}-K_1}+K_7\right]\frac{\Delta T_o}{A_o}+\left[\frac{K_4K_8}{\frac{Mh_o}{F}-K_1}+K_9\right]\frac{\Delta h_t}{h_t}+\frac{K_8M/F}{\frac{Mh_o}{F}-K_1}\Delta S \quad (22)$$

$$\frac{\Delta h_m}{h_m}=\left[\frac{K_2K_9}{\frac{Mh_o}{F}-K_1}+K_{10}\right]\frac{\Delta T_t}{A_t}+\left[\frac{K_3K_9}{\frac{Mh_o}{F}-K_1}+K_{11}\right]\frac{\Delta T_o}{A_t}+\left[\frac{K_4K_9}{\frac{Mh_o}{F}-K_1}+K_{12}\right]\frac{\Delta h_t}{h_t}+\frac{K_9M/F}{\frac{Mh_o}{F}-K_1}\Delta S \quad (23)$$

In the roll gap of a particular stand, a mass flow relationship may be written

$$h_tv_t=h_mv_m=h_ov_o \quad (24)$$

Density is constant through the roll gap because plastic flow does not produce a change in density.[3] The width of the strip remains the same. Thus, density and width do not appear in equation 24.

If the analysis is restricted from this point on to small changes in speed, it will be possible to develop linear equations far simpler to analyze than would be otherwise possible. Replace each of the variables in equation 24 by an initial value plus a change

$$\left.\begin{aligned}h_t&=h_t+\Delta h_t\\h_m&=h_m+\Delta h_m\\h_o&=h_o+\Delta h_o\\v_t&=v_t+\Delta v_t\\v_m&=v_m+\Delta v_m\\v_o&=v_o+\Delta v_o\end{aligned}\right\} \quad (25)$$

with the reference point such that

$$h_tv_t=h_mv_m=h_ov_o \quad (26)$$

Substituting equations 25 into equation 24 and dividing by equation 26

$$\left(1+\frac{\Delta h_t}{h_t}\right)\left(1+\frac{\Delta v_t}{v_t}\right)=\left(1+\frac{\Delta h_m}{h_m}\right)\left(1+\frac{\Delta v_m}{v_m}\right) \quad (27)$$

$$\left(1+\frac{\Delta h_o}{h_o}\right)\left(1+\frac{\Delta v_o}{v_o}\right)=\left(1+\frac{\Delta h_m}{h_m}\right)\left(1+\frac{\Delta v_m}{v_m}\right) \quad (28)$$

Since the delta quantities are small compared to unity, the products of two delta quantities may be neglected to yield

$$\frac{\Delta v_t}{v_t}=\frac{\Delta v_m}{v_m}+\frac{\Delta h_m}{h_m}-\frac{\Delta h_t}{h_t} \quad (29)$$

$$\frac{\Delta v_o}{v_o}=\frac{\Delta v_m}{v_m}+\frac{\Delta h_m}{h_m}-\frac{\Delta h_o}{h_o} \quad (30)$$

Equations 21 and 23 may be substituted into equations 29 and 30 to yield

$$\frac{\Delta v_t}{v_t}=\frac{\Delta v_m}{v_m}+\left[\frac{K_2K_9}{\frac{Mh_o}{F}-K_1}+K_{10}\right]\frac{\Delta T_t}{A_t}+\left[\frac{K_3K_9}{\frac{Mh_o}{F}-K_1}+K_{11}\right]\frac{\Delta T_o}{A_o}+\left[\frac{K_4K_9}{\frac{Mh_o}{F}-K_1}+K_{12}-1\right]\frac{\Delta h_t}{h_t}+\frac{K_9\frac{M}{F}}{\frac{Mh_o}{F}-K_1}\Delta S \quad (31)$$

$$\frac{\Delta v_o}{v_o}=\frac{\Delta v_m}{v_m}+\left[\frac{K_2(K_9-1)}{\frac{Mh_o}{F}-K_1}+K_{10}\right]\frac{\Delta T_t}{A_t}+\left[\frac{K_3(K_9-1)}{\frac{Mh_o}{F}-K_1}+K_{11}\right]\frac{\Delta T_o}{A_o}+\left[\frac{K_4(K_9-1)}{\frac{Mh_o}{F}-K_1}+K_{12}\right]\frac{\Delta h_t}{h_t}+\frac{(K_9-1)\frac{M}{F}}{\frac{Mh_o}{F}-K_1}\Delta S \quad (32)$$

The equations describing plastic flow in the roll gap are now in the form required for substitution into equations 1 through 5. Equations 1 through 5 must be written in small oscillation form.

$$p\frac{\Delta T_1}{A_1}=\frac{Ev_{10}}{L_4}\left[\frac{\Delta v_{1f}}{v_{1f}}-\frac{\Delta v_{10}}{v_{10}}\right] \quad (33)$$

$$p\frac{\Delta T_2}{A_2}=\frac{Ev_{20}}{L_4}\left[\frac{\Delta v_{1f}}{v_{1f}}-\frac{\Delta v_{20}}{v_{20}}\right] \quad (34)$$

$$p\frac{\Delta T_3}{A_3}=\frac{Ev_{30}}{L_4}\left[\frac{\Delta v_{1f}}{v_{1f}}-\frac{\Delta v_{30}}{v_{30}}\right] \quad (35)$$

$$p\frac{\Delta T_4}{A_4}=\frac{Ev_{40}}{L_4}\left[\frac{\Delta v_{1f}}{v_{1f}}-\frac{\Delta v_{40}}{v_{40}}\right] \quad (36)$$

$$p\frac{\Delta T_5}{A_5}=\frac{Ev_{50}}{L_4}\left[\frac{\Delta v_R}{v_R}-\frac{\Delta v_{50}}{v_{50}}\right] \quad (37)$$

where T_1/A_1, T_2/A_2, T_3/A_3, T_4/A_4, and T_5/A_5 are the output tensions of stands *1* through *5*.

As pointed out in Appendix I, the output tension of one stand in pounds is equal to the input tension to the next. Since variations in gauge are small in a well-constructed mill with a good control, little error will result from assuming that output tension of a stand in pounds per square inch is equal to input tension of the next stand in pounds per square inch. A simplification of equations 31 and 32 results.

The purpose of the steps which follow is to get a set of equations analogous to

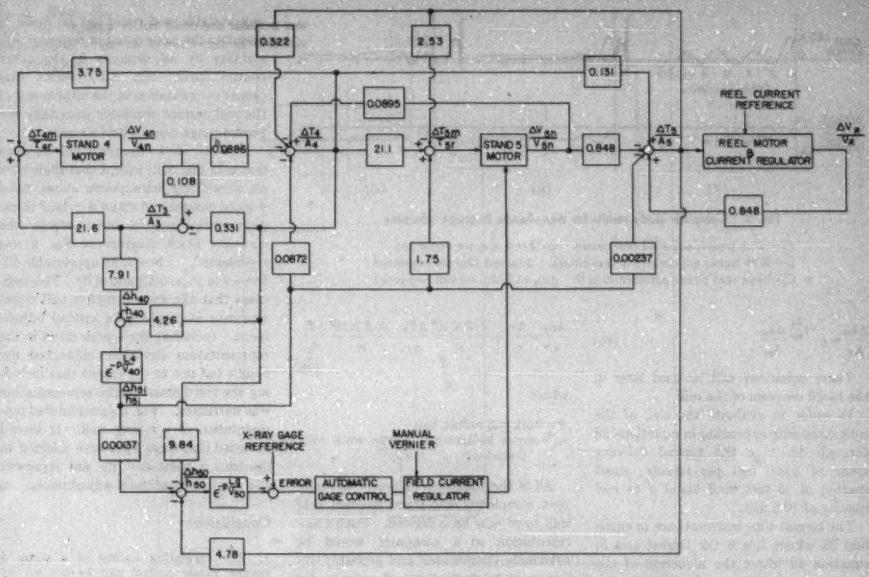


Fig. 4. Block diagram of continuous gauge control including stands 4, 5, and winding reel

equations 11 through 16 which were derived assuming no reduction was being taken. Much can be gained at this point if numerical values for each stand from Table I are substituted into equations 31 and 32. These in turn may be substituted into equations 33 through 37. A typical value of spring constant M is 40 million pounds per inch. The modulus of elasticity for cold steel is 30 million pounds per square inch. The strip width is 36 inches. The result is

$$\frac{\Delta T_1}{A_1}=\frac{0.125}{1+p\,0.00415\frac{L_1}{v_{10}}}\left[-\frac{\Delta v_{1n}}{v_{1n}}+\frac{\Delta v_{2n}}{v_{2n}}-\right.$$
$$1.45\,\frac{\Delta T_2}{A_2}-0.0174\,\frac{\Delta h_{1f}}{h_{1f}}-0.583\,\frac{\Delta h_{2f}}{h_{2f}}+$$
$$\left.0.551\Delta S_1+7.23\Delta S_2\right] \quad (38)$$

$$\frac{\Delta T_2}{A_2}=\frac{0.109}{1+p\,0.00363\frac{L_2}{v_{20}}}\left[-\frac{\Delta v_{2n}}{v_{2n}}+\frac{\Delta v_{3n}}{v_{3n}}+\right.$$
$$0.818\,\frac{\Delta T_1}{A_1}-2.28\,\frac{\Delta T_3}{A_3}-0.0119\,\frac{\Delta h_{2f}}{h_{2f}}-$$
$$\left.0.573\,\frac{\Delta h_{3f}}{h_{3f}}+0.666\Delta S_2+8.63\Delta S_3\right] \quad (39)$$

$$\frac{\Delta T_3}{A_3}=\frac{0.109}{1+p\,0.00363\frac{L_3}{v_{30}}}\left[-\frac{\Delta v_{3n}}{v_{3n}}+\frac{\Delta v_{4n}}{v_{4n}}+\right.$$
$$0.451\,\frac{\Delta T_2}{A_2}-3.08\,\frac{\Delta T_4}{A_4}-0.00961\,\frac{\Delta h_{3f}}{h_{3f}}-$$
$$\left.0.668\,\frac{\Delta h_{4f}}{h_{4f}}+0.861\Delta S_3+8.17\Delta S_4\right] \quad (40)$$

$$\frac{\Delta T_4}{A_4}=\frac{0.0895}{1+p\,0.00298\frac{L_4}{v_{40}}}\left[-\frac{\Delta v_{4n}}{v_{4n}}-\frac{\nabla v_{5n}}{v_{5n}}+\right.$$
$$0.101\,\frac{\Delta T_3}{A_3}-3.60\,\frac{\Delta T_5}{A_5}-0.0287\,\frac{\Delta h_{4f}}{h_{4f}}-$$
$$\left.0.974\,\frac{\Delta h_{5f}}{h_{5f}}+0.990\Delta S_4+9.85\Delta S_5\right] \quad (41)$$

$$\frac{\Delta T_5}{A_5}=\frac{0.848}{1+p\,0.0282\frac{L_5}{v_{50}}}\left[-\frac{\Delta v_{5n}}{v_{5n}}+\frac{\Delta v_R}{v_R}+\right.$$
$$\left.0.154\,\frac{\Delta T_4}{A_4}-0.0279\,\frac{\Delta h_{5f}}{h_{5f}}+0.989\Delta S_5\right] \quad (42)$$

Equations 38 through 42, though similar in form to the elastic equations 12 through 16, are more complicated. The coefficient of the bracket in equations 38 through 41 is about 10 per cent of the corresponding coefficient of equation 42. This coefficient relates tension to speed. When tensions T_1, T_2, T_3, and T_4 tend to increase due to speed changes, they are relieved by the plasticity of the steel in both the preceding and the succeeding roll gap. The latter effect is the more pronounced. When tension T_5 tends to increase, it is relieved only by the preceding roll gap since the reel does not take reduction. Hence the higher "gain" in equation 42. In addition, the reduction of the time constant in equation 42 is less by the same ratio.

The quantity Δh_{1f} in equation 38 is simply the variation in the incoming strip thickness. A variation in the thickness of material delivered from one stand will arrive at the next stand at a later time and appear as a change in input thickness. The time involved in this propagation is simply the stand spacing divided by the interstand strip velocity. The equations which express this relationship are

$$\frac{\Delta h_{2f}}{h_{2f}}=e^{-p\frac{L_1}{v_{10}}}\frac{\Delta h_{1o}}{h_{1o}} \quad (43)$$

$$\frac{\Delta h_{3f}}{h_{3f}}=e^{-p\frac{L_2}{v_{20}}}\frac{\Delta h_{2o}}{h_{2o}} \quad (44)$$

$$\frac{\Delta h_{4f}}{h_{4f}}=e^{-p\frac{L_3}{v_{30}}}\frac{\Delta h_{3o}}{h_{3o}} \quad (45)$$

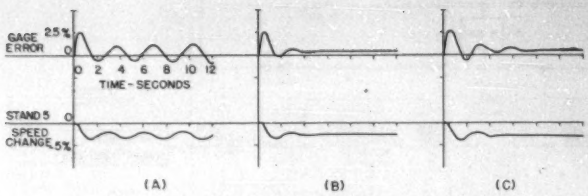


Fig. 5. Computer study results for step change in gauge reference

A—With poorly adjusted reel circuit. $\Delta v_{\cdot a}$ and Δh_{5i} are neglected
B—With better adjustments in reel circuit. Δv_{5a} and Δh_{5i} are neglected
C—Same reel circuit adjustments as B. Δv_{5a} and Δh_{5i} are not neglected

$$\frac{\Delta h_{5i}}{h_{5i}} = {}_{4}\!-\gamma_{5a}\frac{L_4}{v_{4a}}\frac{\Delta h_{4o}}{h_{4o}} \tag{46}$$

These equations will be used later in the block diagram of the mill.

In order to evaluate the size of the time constants appearing in equations 38 through 42, take the typical delivery speed of 3,600 feet per minute, stand spacing of 13 feet, and stand *5* to reel spacing of 10.5 feet.

The largest time constants are in equation 38 where L/v is the largest and in equation 42 where the reduction of the time constant due to plasticity in the roll gap is the least. They are, respectively, 0.0071 second and 0.0062 second. Since they are so short, any analysis of a control system may logically ignore them. Thus, the equations become algebraic rather than differential equations yet still describe the transient behavior of the mill. It is thus seen that disturbances travel at two different speeds in a tandem mill. Tension disturbances are felt almost instantly throughout the mill. Thickness variations move much more slowly.

Equations have thus been derived which relate delivered gauge to tension, incoming gauge, and screw changes (equation 21); rolling torque to the same quantities (equation 22); and tensions to speed changes, incoming thickness, and screw changes (equations 38 through 42). Stand speed is a function of motor load torque so, to be useful, the rolling torque equations should be written to give motor load torques.

An individual stand motor must supply the rolling torque plus the torque due to incoming tension minus the torque due to outgoing tension or in small oscillation form.

$$\Delta\tau_m = \Delta\tau_r + A_i R\times 10^6\frac{\Delta T_i}{A_i} - A_o R\times 10^6\frac{\Delta T_o}{A_o} \tag{47}$$

Dividing through by τ_r to put the equation in per-unit form

$$\frac{\Delta\tau_m}{\tau_r} = \frac{\Delta\tau_r}{\tau_r} + \frac{A_i R\times 10^6}{\tau_r}\frac{\Delta T_i}{A_i} - \frac{A_o R\times 10^6}{\tau_r}\frac{\Delta T_o}{A_o} \tag{48}$$

where

R = work roll radius, feet
τ_m = motor load torque at the work rolls, pound-feet

All of the equations necessary to represent completely a 5-stand tandem cold mill have now been derived. Such a representation on a computer would be extremely complicated and probably unnecessary to study the performance of a stand *5* vernier. Appendix III deals with the simplification to an equivalent mill.

Block Diagram for Study of Stand 5 Vernier

Fig. 4 is the block diagram of the equivalent mill, the equations of which were derived in Appendix III, the automatic gauge control, stand *5* field current regulator, the motors of stands *4*, *5*, and the winding reel, and the reel current regulator. The block diagram shows that the presence of strip in the mill results in a considerable number of important mechanical interconnections among stands *4*, *5*, and the reel. In general, all three will have regulating systems of some kind associated with them which must be considered.

To illustrate this point, Fig. 5 shows selected results of an analogue computer study. Fig. 5 shows the response of the system of Fig. 4 to a step change in reference to the X-ray gauge and vernier automatic gauge control. Figs. 5(A) and 5(B) are computer results when stand *4* speed variations Δv_{4a} and stand *5* input thickness variations Δh_{5i} caused by this step change were neglected. The reel current regulator considered by itself was perfectly stable. Fig. 5(A) shows the response of the gauge control system to this step change in reference. The system was unstable and could not be made to maintain the same speed of response with stability by adjustments in the gauge control itself. Fig. 5(B) shows that proper co-ordination of the adjustment of the reel current regulator materially improves gauge control performance.

Fig. 5(C) shows the results of the same test as in Fig. 5(B) except that there were no simplifying assumptions about stand *4* speed change and stand *4* output thickness variations. In other words, the complete block diagram of Fig. 4 was represented. There is an appreciable difference in Figs. 5(C) and 5(B). This indicates that the more complete mill representation is required for critical adjustment. Including the fourth stand in the representation obviously influenced the results but not to the extent that including the third stand in the representation was warranted. Fig. 6 shows limited representation of a rolling mill. It should be noted that these runs were selected to illustrate a point and do not represent the results of optimum adjustments.

Conclusions

1. The regulating ability of a stand *5* vernier gauge control can be studied by considering disturbances originating near the delivery end of the mill only. Since there is a limited tendency for these disturbances to reflect back into the early stands, they do not produce appreciable stand speed or outgoing thickness variations in stands *1*, *2*, and *3*. The behavior of these stands need not then be considered for such studies.

2. Preliminary studies of the stand *5* vernier control may be made assuming that stand *4* speed does not change and that output thickness of stand *4* does not change.

3. The more complete mill representation considering changes in speed and outgoing thickness of stand *4* is required for more accurate analysis.

4. The adjustment of any control systems which influence stand *4* speed, stand *5* speed, and reel speed are an integral part of a vernier gauge control. Therefore, it is not possible to engineer an optimum stand *5* vernier gauge control without knowing the transient characteristics of each regulating system and without having control of its design and adjustment.

Appendix I. Derivation of Interstand Tension Equations

If the equations are derived for stands *4* and *5*, the equations for the remaining stands will be readily apparent.

Fig. 7 shows stand *4*, stand *5*, and the reel with the strip moving from left to right.

Elongation is the increase in length of a particular section when subjected to stress. Since the strip between stands is not, in general, uniform in cross-sectional area and the total pounds of tension are the same for

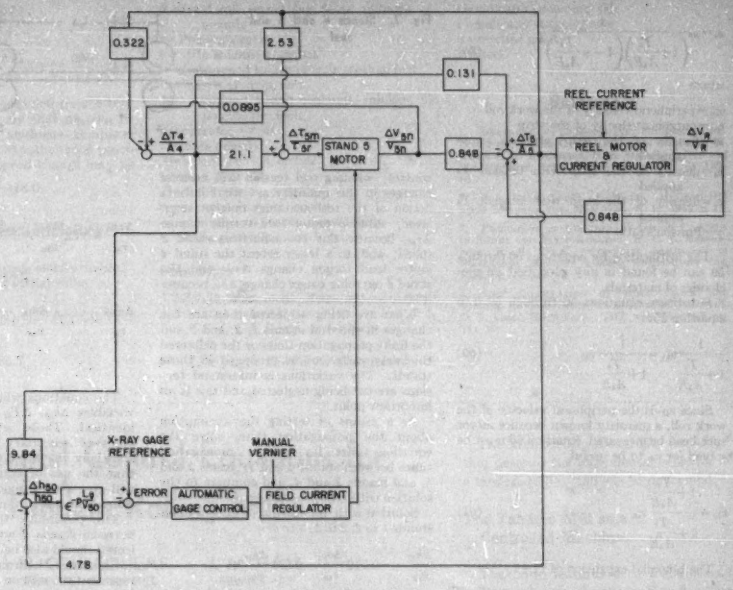


Fig. 6. Block diagram of limited representation of rolling mill

each cross section, the stress is not the same for each cross section. The variations in cross section are due to the fact that no mill can roll strip to absolutely uniform cross section. However, a well-controlled mill will inherently roll strip to close tolerance. Therefore, assume that stress between stands is equal to total tension divided by the nominal cross-sectional area. Let

L_4 = the interstand spacing (also the length of strip between stands measured under stress), feet
L_R = distance from stand 5 to the reel, feet
l_4 = unstretched length of strip between stands 4 and 5, feet
e_4 = total elongation between stands 4 and 5 feet
A_4 = nominal cross-sectional area between stands 4 and 5, square inches
T_4 = total pounds tension between stands 4 and 5, millions of pounds
E = modulus of elasticity, millions of psi

Since the strip is elastic between stands, it obeys Hooke's law and

$$e_4 = \frac{T_4 l_4}{A_4 E} \tag{49}$$

From the way in which the quantities have been defined

$$e_4 + l_4 = L_4 \tag{50}$$

When tension is applied to steel between stands in a cold reduction mill, the maximum practical value the resultant stress can reach is 60,000 pounds per square inch. Since the modulus of elasticity for steel is 30 million pounds per square inch, this would produce an elongation of only 0.2 per cent, which may be written

$$\frac{e_4}{l_4} = 0.002 \tag{51}$$

From an examination of equation 49 it is seen that elongation e_4 is proportional to the product of tension T_4 and the unstretched length l_4. It can be seen from equations 49, 50, and 51 that the maximum error in computation of elongation that might result from using L_4 in place of l_4 is 0.2 per cent. Therefore, to a very good approximation, equation 49 may be written

$$e_4 = \frac{T_4 L_4}{A_4 E} \tag{52}$$

As was stated before, elongation is the increase in length of a particular section under stress. A change in elongation between stands is the consequence of a difference in velocity of the strip leaving one stand and that entering the next, or

$$e_4 = \int_o^t [v_{5l} - v_{4e}]\,dt \tag{53}$$

where

v_{4e} = velocity of the strip as it leaves stand 4, feet per second
v_{5l} = velocity of the strip as it enters stand 5, feet per second
t = time, seconds

It should be appreciated at this point that elongation e_4 now has a definition in time. That is, e_4 may well have a value at time zero.

Equation 52 may be equated to 53:

$$\frac{T_4}{A_4} = \frac{E}{L_4}\int_o^t [v_{5l} - v_{4e}]\,dt \tag{54}$$

Appendix II. Derivation of Roll Gap Equations with No Deformation in the Roll Gap

Consider that the strip is entering stand 1 at constant area, A_0. Since it has no tension on it, it is the unstressed area out of each of the stands. Assume that there is but one point of contact between the work roll and the strip and that there is no slippage between them. (Actually the material is deformed elastically but this is a fair approximation sufficient for illustration of a principle.) The entering velocity of the strip and the work roll peripheral velocity will be the same. There will be a discontinuity at the point of contact if the incoming tension and outgoing tension are not equal. Across this boundary there must be a continuity of mass flow because there is no storage volume. Then

$$\rho_3 A_3 v_{3e} = \rho_4 A_4 v_{4e} \tag{55}$$

$$A_3 = A_0\left(1 - \gamma\frac{T_3}{A_3 E}\right)^2 \tag{56}$$

$$A_4 = A_0\left(1 - \gamma\frac{T_4}{A_4 E}\right)^2 \tag{57}$$

$$\rho_3 = \rho_0\frac{1}{\left(1 + \frac{T_3}{A_3 E}\right)\left(1 - \gamma\frac{T_3}{A_3 E}\right)^2} \tag{58}$$

$$\rho_t = \rho_0 \frac{1}{\left(1+\frac{T_t}{A_tE}\right)\left(1-\gamma\frac{T_t}{A_tE}\right)^2} \quad (59)$$

where

v_{tn} = peripheral velocity of the work roll
v_{to} = outgoing velocity of the strip
ρ_0 = density of the strip with no stress applied
ρ_2 = density of the strip with tension T_2 applied
ρ_t = density of the strip with tension T_t applied
γ = Poisson's ratio

The justification for equations 56 through 59 can be found in any good text on mechanics of materials.

Substitute equations 56 through 59 into equation 55:

$$\frac{1}{1+\frac{T_2}{A_2E}} v_{2n} = \frac{1}{1+\frac{T_t}{A_tE}} v_{to} \quad (60)$$

Since v_{tn} is the peripheral velocity of the work roll, a quantity known because motor speed can be measured, equation 60 must be solved for v_{to} to be useful.

$$v_{to} = \frac{1+\frac{T_t}{A_tE}}{1+\frac{T_2}{A_2E}} v_{tn} \quad (61)$$

The binomial expansion of $(1+x)^{-1}$ is

$$(1+x)^{-1} = 1 - \frac{x}{1!} + \frac{x^2}{2!} - \frac{x^3}{3!} + \frac{x^4}{4!} - \ldots \quad (62)$$

If x is much less than unity

$$(1+x)^{-1} \cong 1 - x \quad (63)$$

Using the approximate equation 63 in equation 61

$$v_{to} \cong \left(1+\frac{T_t}{A_tE}\right)\left(1-\frac{T_2}{A_2E}\right) v_{tn} \quad (64)$$

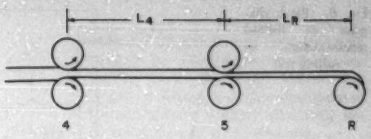
Since the quantity T_t/A_tE is per-unit strain, and was shown by equation 51 to be less than 0.2 per cent, little error in the calculation of velocity will result if the product of the strains is neglected. Thus

$$v_{to} \cong \left(1+\frac{T_t}{A_tE}-\frac{T_2}{A_2E}\right) v_{tn} \quad (65)$$

Appendix III. Derivation of Limited Computer Representation

Since the vernier control works at the delivery end of the mill, it is unlikely that the speed changes there will produce appreciable disturbances at the entry end of the mill. The physical picture of the process involved indicates that if stand *5* speed is controlled by the thickness vernier, the controlled quantities of most importance are: stand *5* speed change Δv_{5n}, reel speed change Δv_R, and to a lesser extent upon the change of speed of stand *4* Δv_{4n} due to load change $\Delta\tau_{4m}$. The output quantities which must be represented are stand *5* delivered gauge Δh_{5o} since this is what initiates corrective action in the automatic gauge control; winding reel tension ΔT_5 because changes in this quantity are what initiate action of the reel armature current regulator; stand *5* motor load torque change $\Delta\tau_{5m}$ because this too influences stand *5* speed; and, to a lesser extent the stand *4* motor load torque change $\Delta\tau_{4m}$ and the stand *4* outgoing gauge change Δh_{4o} because it later becomes an input to stand *5*.

Fig. 7. Stands 4 and 5 and reel



What are being neglected then are the changes in speed of stands *1*, *2*, and *3* and the finite propagation times of the delivered thickness variations as produced in those stands. The variations in interstand tensions are not being neglected and this is an important point.

As a means of testing the assumption about the propagation time, solve the equations first for infinite propagation times between stands *1* and *2*, stands *2* and *3*, and stands *3* and *4*, and compare to the solution with zero propagation times.

Solution with infinite propagation times stands *1* to *2*, *2* to *3*, *3* to *4*:

$$\frac{\Delta h_{4o}}{h_{4o}} = -0.710\frac{\Delta v_{4n}}{v_{4n}} - 0.571\frac{\Delta v_{5n}}{v_{5n}} + 0.431\frac{\Delta v_R}{v_R} + 0.125\frac{\Delta h_{5t}}{h_{5t}} \quad (66)$$

$$\frac{\Delta h_{5o}}{h_{5o}} = 0.887\frac{\Delta v_{4n}}{v_{4n}} + 0.421\frac{\Delta v_{5n}}{v_{5n}} - 1.32\frac{\Delta v_R}{v_R} + 0.908\frac{\Delta h_{5t}}{h_{5t}} \quad (67)$$

$$\frac{\Delta T_5}{A_5} = -0.0111\frac{\Delta v_{4n}}{v_{4n}} - 0.803\frac{\Delta v_{5n}}{v_{5n}} + 0.814\frac{\Delta v_R}{v_R} - 0.0336\frac{\Delta h_{5t}}{h_{5t}} \quad (68)$$

$$\frac{\Delta\tau_{4m}}{\tau_{4r}} = 3.23\frac{\Delta v_{4n}}{v_{4n}} - 3.79\frac{\Delta v_{5n}}{v_{5n}} + 2.85\frac{\Delta v_R}{v_R} + 0.829\frac{\Delta h_{5t}}{h_{5t}} \quad (69)$$

$$\frac{\Delta\tau_{5m}}{\tau_{5r}} = -1.76\frac{\Delta v_{4n}}{v_{4n}} + 9.37\frac{\Delta v_{5n}}{v_{5n}} - 7.59\frac{\Delta v_R}{v_R} + 0.225\frac{\Delta h_{5t}}{h_{5t}} \quad (70)$$

Solution with zero propagation times stands *1* to *2*, *2* to *3*, *3* to *4*:

$$\frac{\Delta h_{4o}}{h_{4o}} = -0.885\frac{\Delta v_{4n}}{v_{4n}} - 0.422\frac{\Delta v_{5n}}{v_{5n}} + 0.318\frac{\Delta v_R}{v_R} + 0.0924\frac{\Delta h_{5t}}{h_{5t}} \quad (71)$$

$$\frac{\Delta h_{5o}}{h_{5o}} = 0.880\frac{\Delta v_{4n}}{v_{4n}} + 0.426\frac{\Delta v_{5n}}{v_{5n}} - 1.32\frac{\Delta v_R}{v_R} + 0.907\frac{\Delta h_{5t}}{h_{5t}} \quad (72)$$

$$\frac{\Delta T_5}{A_5} = -0.0110\frac{\Delta v_{4n}}{v_{4n}} - 0.803\frac{\Delta v_{5n}}{v_{5n}} + 0.814\frac{\Delta v_R}{v_R} - 0.0336\frac{\Delta h_{5t}}{h_{5t}} \quad (73)$$

$$\frac{\Delta\tau_{4m}}{\tau_{4r}} = 3.22\frac{\Delta v_{4n}}{v_{4n}} - 3.77\frac{\Delta v_{5n}}{v_{5n}} + 2.84\frac{\Delta v_R}{v_R} + 0.826\frac{\Delta h_{5t}}{h_{5t}} \quad (74)$$

$$\frac{\Delta\tau_{5m}}{\tau_{5r}} = -1.75\frac{\Delta v_{4n}}{v_{4n}} + 9.36\frac{\Delta v_{5n}}{v_{5n}} - 7.58\frac{\Delta v_R}{v_R} + 0.227\frac{\Delta h_{5t}}{h_{5t}} \quad (75)$$

The equations which define the primary variables Δh_{5o}, ΔT_5, and $\Delta\tau_{5m}$ are almost identical. Those which define the variables of secondary importance are substantially the same. It can be concluded that the use of equations 71 through 75 will give a very good transient representation of the tandem mill.

The possibility that propagation time between stands *4* and *5* may also be neglected should also be investigated.

Equations 71 through 75 could be used to represent the mill on an analogue computer. However, interstand tension has been eliminated as a variable and the physical picture says that delivered thickness from a stand is dependent on both incoming and outgoing tensions. If a good physical representation, rather than a purely mathematical one, is desired, use should be made of the foregoing information to achieve this end.

It has been discovered here that it is unnecessary to consider variation in speed of the first three stands, v_{1n}, v_{2n}, and v_{3n}. In addition, the fact that propagation times *1* to *2*, *2* to *3*, and *3* to *4* may be assumed infinite really means that input thickness changes Δh_{2t}, Δh_{3t}, and Δh_{4t} may be neglected. Screw changes may be neglected since they are not being considered. With these simplifying assumptions, equations 21, 40, 41, 42, and 48 become

$$\frac{\Delta h_{4o}}{h_{4o}} = -7.91\frac{\Delta T_3}{A_3} - 4.26\frac{\Delta T_4}{A_4} \quad (76)$$

$$\frac{\Delta h_{5o}}{h_{5o}} = -9.84\frac{\Delta T_4}{A_4} - 4.78\frac{\Delta T_5}{A_5} - 0.00137\frac{\Delta h_{5t}}{h_{5t}} \quad (77)$$

$$\frac{\Delta T_2}{A_2} = 0.108\frac{\Delta v_{1n}}{v_{1n}} - 0.331\frac{\Delta T_4}{A_4} \quad (78)$$

$$\frac{\Delta T_4}{A_4} = -0.0886\frac{\Delta v_{3n}}{v_{3n}} + 0.0895\frac{\Delta v_{4n}}{v_{4n}} - 0.322\frac{\Delta T_3}{A_3} - 0.0872\frac{\Delta h_{5t}}{h_{5t}} \quad (79)$$

$$\frac{\Delta T_5}{A_5} = -0.848\frac{\Delta v_{5n}}{v_{5n}} + 0.848\frac{\Delta v_R}{v_R} + 0.131\frac{\Delta T_4}{A_4} - 0.0237\frac{\Delta h_{5t}}{h_{5t}} \quad (80)$$

$$\frac{\Delta\tau_{3m}}{\tau_w} = 21.6\frac{\Delta T_2}{A_2} - 3.75\frac{\Delta T_3}{A_3} \quad (81)$$

$$\frac{\Delta\tau_{4m}}{\tau_w} = 21.1\frac{\Delta T_4}{A_4} - 2.53\frac{\Delta T_3}{A_3} + 1.75\frac{\Delta h_{5i}}{h_{5i}} \quad (82)$$

To test whether propagation from 4 to 5 is important requires only that $\Delta h_{5i}/h_{5i}$ be assumed zero in an analogue computer study for comparison. Another good possibility is that stand 4 speed change may be neglected.

Summary of Nomenclature

A = strip cross-section area, square inches
Δ = denotes change from initial value
e = elongation, feet
E = modulus of elasticity, millions of pounds per square inch
F = total roll separating force, millions of pounds
ν = Poisson's ratio
h = strip thickness, inches
l = unstretched length of strip section, feet
L = length of strip, feet
M = combined spring constant, millions of pounds per inch
p = operator = d/dt
ρ = mass density
S = screw position, inches
t = time, seconds
T = total tension, pounds
v = velocity, feet per second
τ_m = motor load torque, pound-feet
τ_r = rolling torque, pound-feet

Subscripts

1,2,3,4,5 = when used alone indicate nominal values between stands *1–2*, *2–3*, *3–4*, *4–5*, *5*–reel respectively
i = indicates incoming side
o = indicates outgoing side
n = neutral point
R = reel

(*See pages 370–74 for discussion.*)

References

1. Development of Automatic Thickness Control for Tandem Cold Reduction Mills, R. W. Holman, M. W. Cannon, H. S. Marwell. *Iron and Steel Engineer*, Pittsburgh, Pa., Aug. 1956, pp. 132–40.

2. Plasticity (book), A. Nadai. McGraw-Hill Book Company, Inc., New York, N. Y., 1943, p. 78.

3. Principles of Continuous Gauge Control in Sheet and Strip Rolling, W. C. F. Hessenberg, R. B. Sims. *Proceedings*, Institution of Mechanical Engineers, London, England, vol. 166, 1952, pp. 75–81.

4. The Calculation of Roll Force and Torque in Cold Strip Rolling with Tensions, D. R. Bland, Hugh Ford. *Ibid.*, vol. 159, 1948, pp. 144–53.

Incremental Control Equations for Tandem Rolling Mills

J. H. COURCOULAS
ASSOCIATE MEMBER AIEE

J. M. HAM
MEMBER AIEE

Synopsis: This study illustrates the nature of the multivariable control problem in a particular context. The character of transfer or control equations for multivariable machines and processes is briefly examined. Incremental steady-state transfer equations are developed for a typical multivariable machine, namely, the strip rolling mill. The equations are derived from the theory of plastic deformation in rolling, and the electromechanical equilibrium equations of a mill. A numerical example of these equations for a representative 5-stand mill is given. The results may be used to consider general forms for automatic control systems.

ANALYTICAL methods for the design of single-variable control systems having a single primary input and a corresponding output have been well developed.[1–5] These design procedures require knowledge of the dynamic relationship between the input to, and the response from, the controlled element. Suitable transfer or describing functions are known for most single-variable systems. However, for multivariable control systems, analytical design methods are only now in the course of development[6–8] and much knowledge remains to be gained about the transfer functions for multivariable controlled elements such as chemical processes, rolling mills for paper and metals, etc.

This paper describes the development of transfer functions for tandem rolling mills used in the manufacture of thin metal strips and sheets. It is shown that knowledge of the transfer functions points directly to preferred forms of automatic control.

A block diagram for a multivariable control system is shown in Fig. 1. This system has n inputs, m outputs, k control variables, and j disturbance variables. The box labelled "computer" performs the operation of comparing the output signals with the desired output signals and adjusting or compensating the results of this comparison so that the signal z_i will lead to an optimum response y_i. The linear performance of such a system is conveniently described by equations of the form

$$Y_i = \sum_r H_{ir} Z_r + \sum_r H_{ir}^* U_r \quad (i = 1,2 \ldots m) \quad (1)$$

where

U_r = Laplace transform of the disturbance $u_r(t)$
Z_r = Laplace transform of the control signal $z_r(t)$
H_{ir} = Laplace transfer function relating $y_i(t)$ to $z_r(t)$
H_{ir}^* = Laplace transfer function relating $y_i(t)$ to $u_r(t)$

The transfer functions H_{ir} and H_{ir}^* that describe the incremental behavior of a tandem rolling mill will be developed.

The Tandem Mill as a Controlled Machine

The basic features at a 5-stand cold rolling mill are shown in Fig. 2. The material enters stand *1* at a thickness of h_0 and leaves the mill from stand *5* at a thickness of h_5. In each stand the roll force F applied to the material and the torque τ driving the rolls can be varied to control the mill. In addition, the tensions t_0 and t_5 are controllable. The input variables are

h_0 = thickness of input strip
F_i = roll force in stand i $(i = 1,2 \ldots 5)$
τ_i = roll torque in stand i $(i = 1,2 \ldots 5)$
t_0 = input strip tension
t_5 = output strip tension

The significant response variables are

h_5 = output strip thickness
h_{ij} = thickness of strip between stands i and j
t_{ij} = tension in strip between stands i and j
V_{ij} = linear speed of strip between stands i and j
V_5 = speed of output strip

Paper 56-656, recommended by the AIEE Feedback Control Systems Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Summer and Pacific General Meeting, San Francisco, Calif., June 25–29, 1956. Manuscript submitted March 19, 1956; made available for printing April 23, 1956.

J. H. Courcoulas is with Canadian Westinghouse Company, Hamilton, Ont., Canada, and J. M. Ham is with the University of Toronto, Toronto, Ont., Canada.

The authors are grateful to Dr. Hugh Ford and G. Lianis, Department of Mechanical Engineering, Imperial College, London, England, for permission to use unpublished nomograms for the cold-rolling problem. This research was supported in part by the Defence Research Board of Canada under Grant no. *DRB 9001–02*.

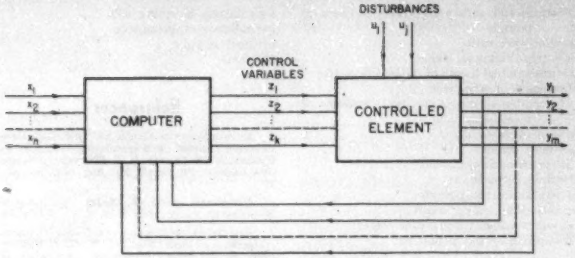


Fig. 1. The basic multivariable control system

The operating objective for the mill is to produce the maximum tonnage of output strip having a gauge h_2 lying within a desired tolerance of a standard thickness h_2^*. Optimum control for a mill, whether it be by human operators or by completely automatic feedback control systems, requires for its development quantitative relations between the input and response variables for the mill. It appears that only recently has the problem of determining the necessary equations attracted much attention in the literature.[9] Control equations for a 5-stand mill are derived in the following section.

Derivation of Incremental Control Equations

In this study, relations are developed among increments in control and response variables measured from given operation points for these variables.

In Fig. 3, one stand of a mill is examined in detail. The material is reduced from a thickness h_1 to h_2 and the two rolls of diameter R are pressed together with a force F that depends on the pressure distribution p along the arc of contact ϕ. The rolls are driven by a torque τ and under load the effective roll radius becomes R', where $R'>R$. The coefficient of friction between rolls and material is μ. There is also a front tension t_2 and a back tension t_1 applied to the strip, which has a yield stress k.

Theories of plastic deformation in rolling[10,11] show that the roll force F can be expressed in general form as

$$F=f(R', p, \phi, \mu, t_2, t_1, h_1, h_2, k) \quad (2)$$

The roll torque can be expressed as the following

$$\tau=g(R, R', p, \phi, \mu, t_1, t_2, h_1, h_2, k) \quad (3)$$

where both functions g and f are nonlinear in their variables.

For a 5-stand mill there will be five sets of equations of the form given by equations 2 and 3. To determine input-output relationships, it is necessary to obtain the simultaneous solution of these equations. As the front tension of one stand is the back tension of the next one and the output gauge of one stand is the input gauge of the next one, and furthermore the volume of material processed per unit time in any stand must be the same, the solution of the equations must satisfy these constraints. The roll torque and roll force equations are nonlinear and therefore the behavior of the mill under general conditions is not readily deduced.

However, it is quite feasible to determine quantitative relationships between small changes in the input and output variables measured from a given set of operating values. The total differentials ΔF and $\Delta\tau$ in equations 3 and 4 are linear in their variables. Hence to find the incremental input-output equations referred to a given schedule of operating conditions, it is only necessary to evaluate ΔF and $\Delta\tau$, introduce the corresponding constraints, and solve the resulting linear equations.

Roll Force versus Thickness Diagram

The determination of the relationship between the roll force F, the strip tensions, strip speeds, and thickness in each stand is now considered.

The mill frame and the rolls themselves are not perfectly rigid. As load is applied on the strip, elastic deformations occur in the mill housing and rolls. Thus the whole structure can be considered as a stiff spring and the displacement under load of the rolls is a measure of the spring compression. Hence the roll force F is given by[9]

$$F=M\Delta S_0 \quad (4)$$

where

M = the spring constant of the stand
ΔS_0 = change in the roll separation under load

Equation 4 can be rewritten as

$$F=M(S-S_0) \quad (5)$$

where

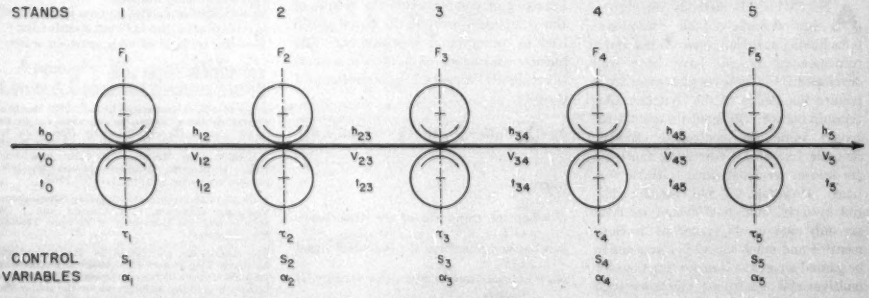


Fig. 2. A 5-stand tandem reduction mill

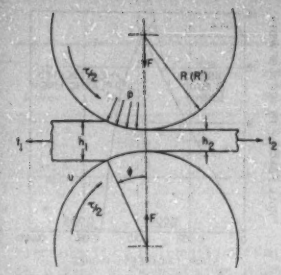


Fig. 3. Variables in the roll gap

S_0 = separation of working roll surfaces when $F=0$

S = separation of rolls during rolling

If the elastic recovery of the strip after passage through the rolls is neglected, the output gauge h_2 in Fig. 2 is equal to S. Under these conditions, equation 5 is written as

$$F = M(h_2 - S_0) \tag{6}$$

and it is called the elastic equation of the stand. When the mill is operating, the roll force F given by equation 6 is equal to the roll force F expressed by equation 2 in terms of strip parameters. Equation 2 is known as the plastic equation and gives a measure of the force required for plastic deformation of the strip. If the input is fixed, the plastic force F can be written

$$F = f(h_2) \tag{7}$$

where F is some function of the output gauge h_2.

The equilibrium condition in the stand is illustrated graphically as point A in Fig. 4. The point A is the point of intersection of the plastic curve $F=f(h_2)$ with the elastic curve for the stand. Point A defines the equilibrium output gauge h_{20}. The output gauge is clearly a function of M, S_0 and of all the variables entering into the plastic force F as defined by equation 2.

Many equations of the form of equations 2 and 3 have been proposed to describe the relations between plastic roll force and the variables in the roll gap of a mill.[10–13] The ones used in this study are those proposed by Ford.[13] Experimental tests have shown that Ford's equations yield suitably accurate results and nomograms have recently been prepared to facilitate use of the equations. If the elastic recovery of the strip is taken in account and the coefficient of friction μ can be determined, the accuracy to be expected is of the order of 2 per cent of measured values.

Roll Torque versus Roll Speed Diagram

During stand operation, the roll torque corresponds to a balance between the torque supplied by the mill motor and the torque required to drive the rolls and deform the strip. It is assumed that the torque supplied by the mill motor is governed by the equation

$$\tau = \alpha + \beta\omega \tag{8}$$

where

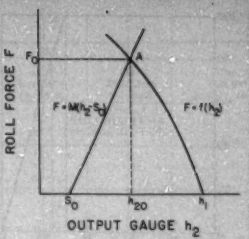


Fig. 4. Roll force versus thickness diagram

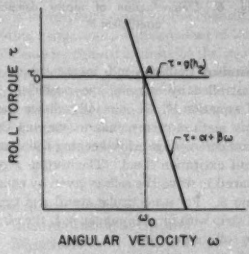


Fig. 5. Roll torque versus angular roll velocity diagram

$\alpha = \tau|_{\omega=0}$ = the stalled torque of the motor

$\beta = \dfrac{\partial\tau}{\partial\omega}$ = the damping coefficient of the drive

ω = angular velocity of the roll shaft

Equation 8 describes drives using d-c shunt motors. In this study it will be

Table I. Incremental Equilibrium Equations for a 5-Stand Mill

	δh_{12}	δh_{22}	δh_{32}	δh_{42}	δh_{52}	$\delta\omega_{12}$	$\delta\omega_{22}$	$\delta\omega_{32}$	$\delta\omega_{42}$	$\delta\omega_{52}$	δh_{13}	δh_{23}	δh_{33}	δh_{43}
$-\left(M_1\delta S_1+\frac{\partial F_1}{\partial\omega_1}\delta\omega_1\right)$	$\frac{\partial F_1}{\partial h_{12}}-M_1$										$\frac{\partial F_1}{\partial h_{13}}$			
$-M_2\delta S_2$	$\frac{\partial F_2}{\partial h_{12}}$	$\frac{\partial F_2}{\partial h_{22}}-M_2$									$\frac{\partial F_2}{\partial h_{13}}$	$\frac{\partial F_2}{\partial h_{23}}$		
$-M_3\delta S_3$		$\frac{\partial F_3}{\partial h_{22}}$	$\frac{\partial F_3}{\partial h_{32}}-M_3$								$\frac{\partial F_3}{\partial h_{13}}$	$\frac{\partial F_3}{\partial h_{23}}$	$\frac{\partial F_3}{\partial h_{33}}$	
$-M_4\delta S_4$			$\frac{\partial F_4}{\partial h_{32}}$	$\frac{\partial F_4}{\partial h_{42}}-M_4$									$\frac{\partial F_4}{\partial h_{33}}$	$\frac{\partial F_4}{\partial h_{43}}$
$-M_5\delta S_5$				$\frac{\partial F_5}{\partial h_{42}}$	$\frac{\partial F_5}{\partial h_{52}}-M_5$									$\frac{\partial F_5}{\partial h_{43}}$
$\delta\omega_1-\frac{\partial\tau_1}{\partial h_1}\delta h_1$	$\frac{\partial\tau_1}{\partial h_{12}}$					$-\beta_1$					$\frac{\partial\tau_1}{\partial h_{13}}$			
$\delta\omega_2$	$\frac{\partial\tau_2}{\partial h_{12}}$	$\frac{\partial\tau_2}{\partial h_{22}}$					$-\beta_2$				$\frac{\partial\tau_2}{\partial h_{13}}$	$\frac{\partial\tau_2}{\partial h_{23}}$	$\frac{\partial\tau_2}{\partial h_{33}}$	
$\delta\omega_3$		$\frac{\partial\tau_3}{\partial h_{22}}$	$\frac{\partial\tau_3}{\partial h_{32}}$					$-\beta_3$				$\frac{\partial\tau_3}{\partial h_{23}}$	$\frac{\partial\tau_3}{\partial h_{33}}$	
$\delta\omega_4$			$\frac{\partial\tau_4}{\partial h_{32}}$	$\frac{\partial\tau_4}{\partial h_{42}}$					$-\beta_4$				$\frac{\partial\tau_4}{\partial h_{33}}$	$\frac{\partial\tau_4}{\partial h_{43}}$
$\delta\omega_5$				$\frac{\partial\tau_5}{\partial h_{42}}$	$\frac{\partial\tau_5}{\partial h_{52}}$					$-\beta_5$				$\frac{\partial\tau_5}{\partial h_{43}}$
0	V_{11}	$-V_{21}$				$h_{12}R_1$	$-h_{22}R_2$							
0		V_{21}	$-V_{31}$				$h_{22}R_2$	$-h_{32}R_3$						
0			V_{31}	$-V_{41}$				$h_{32}R_3$	$-h_{42}R_4$					
0				V_{41}	$-V_{51}$				$h_{42}R_4$	$-h_5R_5$				

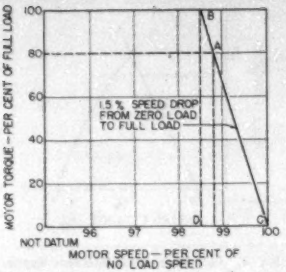


Fig. 6. Determination of motor damping coefficient β

assumed that the mill motor torque is controlled by varying the coefficient α of equation 8. It is readily shown that this method corresponds to varying the armature voltage while keeping the shunt field excitation fixed. The torque τ required to drive the rolls is given by equation 3. For a particular stand and type of strip with fixed input gauge h_1 the plastic roll torque τ is

$$\tau = g(h_2) \tag{9}$$

where τ is some function of the output gauge h_2.

The equilibrium condition in the stand is illustrated graphically as point A in Fig. 5, where the plastic torque line $\tau = g(h_2)$ intersects the mill drive torque line $\tau = \alpha + \beta\omega$. Point A defines the operating speed ω_0 of the rolls. The value of h_2 in the function $\tau = g(h_2)$ is that defined by Fig. 4. It should be noted that τ in equation 9 is not a function of angular roll speed. The roll torque τ associated with the plastic deformation of the strip can be calculated by the equation given by Ford.

Continuity Equations

Equations 6–9 when satisfied simultaneously describe the behavior of a single stand in the mill. In tandem mills adjacent stands are coupled by the strip tension, and during normal operation of the mill, the same volume of material must pass through each stand per unit time. These conditions impose constraints on the variables describing the behavior of the stands. The nature of these constraints is clarified with the help of Fig. 2, which shows the strip and control variables for a 5-stand mill.

In each stand in Fig. 2, there are settings for the parameters S and α that correspond respectively to the separation of the rolls during rolling and to the stalled torque at the roll shaft. The input strip tension t_0 and the output strip tension t_5 will be considered set by reeling systems. With these control settings and any given input gauge h_0 the roll forces F, roll torques τ, strip gauges h, strip tensions t, and strip velocities V will assume definite values.

In view of the continuity condition on the volumetric flow of material through the mill one can write

$$h_{12}V_{12} = h_{23}V_{23} = h_{34}V_{34} = h_{45}V_{45} = h_5V_5 \tag{10}$$

For incremental changes in strip gauges h and strip velocities V, equation 10 becomes

$$h_{12}\delta V_{12} + V_{12}\delta h_{12} = h_{23}\delta V_{23} + V_{23}\delta h_{23} = \ldots = h_5\delta V_5 + V_5\delta h_5 \tag{11}$$

Equation 11 is the differential continuity equation. In this equation the linear strip speeds V are related to the peripheral velocities of the rolls by relations of the form

$$V_2 = v(S_L + 1) \tag{12}$$

where

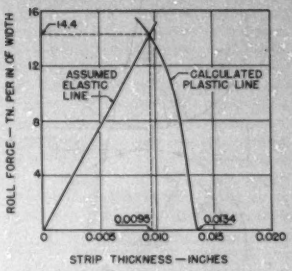


Fig. 7. Estimation of mill stiffness M

S_L = per-unit forward slip in the stand
V_2 = outgoing strip velocity at the stand
v = peripheral velocity of the rolls

It will be assumed that $S_L << 1$ and that $V_2 = v$.

In Fig. 2 the front tension of any stand is equal to the back tension of the following stand. Equations 6–9 for each of the stands, together with the tension coupling constraint and the continuity equation 10, provide a steady-state description of the mill. In the following section these incremental control equations are derived.[14]

Incremental Equilibrium Equations for 5-Stand Mill

The incremental elastic equation for stand 1 obtained from equation 6 is

$$\delta F_1 = M_1(\delta h_{13} - \delta S_1) \tag{13}$$

If the back tension t_0 on stand 1 is fixed, changes in the plastic force of equation 7 are caused by changes in h_0, h_{12}, and t_{12}. Hence the differential of the plastic force can be written as

$$\delta F_1 = \frac{\partial F_1}{\partial h_0}\delta h_0 + \frac{\partial F_1}{\partial h_{12}}\delta h_{12} + \frac{\partial F_1}{\partial t_{12}}\delta t_{12} \tag{14}$$

Table II. Matrix of Equilibrium Equations

	δh_{12}*	δh_{23}	δh_{34}	δh_{45}	δh_5	$\delta\omega_{12}$‡	$\delta\omega_{23}$	$\delta\omega_{34}$	$\delta\omega_{45}$	$\delta\omega_5$	δt_{12}§	δt_{23}	δt_{34}	δt_{45}
$-1,522\delta S_1$*	−2,163										−0.157			
$-1,522\delta S_2$	+183	−1,627	6								−0.985	−0.728		
$-1,522\delta S_3$		+476	−2,104									−1.96	−0.833	
$-1,522\delta S_4$			+1,120	−2,437									−1.00	−0.134
$-1,522\delta S_5$				+542	−3,539									−0.983
$\delta\omega_{12}$†	−510					+402.44					−0.60			
$\delta\omega_{23}$	+307	−92.5					+121.22				+0.455	−0.83		
$\delta\omega_{34}$		+283.5	−359					+27.72				+0.10	−0.20	
$\delta\omega_{45}$			+358	−229					+12.34				+0.175	−0.20
$\delta\omega_5$				+282.5	−328					+5.98				+0.183
0	+60	−120				+0.662	−0.34							
0		+120	−280				+0.34	−0.187						
0			+280	−296				+0.187	−0.141					
0				+296	−420				+0.141	−0.100				

*$\delta h_{12} \ldots \delta h_5$, $\delta S_1 \ldots \delta S_5$, inches
†$\delta\omega_{12}$, $\delta\omega_{23}$, $\delta\omega_{34}$, $\delta\omega_{45}$, $\delta\omega_5$, ton-inches per inch
‡$\delta\omega_{12} \ldots \delta\omega_5$, radians per second
§$\delta t_{12} \ldots \delta t_{45}$, ton per square inch

Table III. 2-Stand Mill Equilibrium Equations

δh_{13}, Inches	δh_{23}, Inches	$\delta\omega_{12}$, Radians per Second	$\delta\omega_{23}$, Radians per Second	δf_{12}, Ton per Square Inch	δS_1, Inches	δS_2, Inches	$\delta\alpha_1$, Ton-Inch per Inch	$\delta\alpha_2$, Ton-Inch per Inch	δf_{12}, Ton per Square Inch
−2,163.0	0	0	0	−0.157	−1,522.0	0	0	0	0
183.0	−1,627.6	0	0	−0.985	0	−1,522.0	0	0	0.728
−510.0	0	402.44	0	−0.600	0	0	1.0	0	0
307.0	−92.5	0	121.22	0.455	0	0	0	1.0	0.830
60.0	−120.0	0.662	−0.340	0	0	0	0	0	0

In equation 14 it is important to note that δh_0, a change in the input gauge, is not controllable and therefore δh_0 should be looked upon as a disturbance to mill operation.

Differential equilibrium in roll force about the operating point given by Fig. 4 is defined by equating the right-hand sides of equation 13 and equation 14 to obtain, on rearrangement of terms,

$$-\left(M_1\delta S_1+\frac{\partial F_1}{\partial h_0}\delta h_0\right)=\left(\frac{\partial F_1}{\partial h_{12}}-M_1\right)\delta h_{12}+\frac{\partial F_1}{\partial f_{12}}\delta f_{12} \quad (15)$$

An equation similar to equation 15 is obtained for stands *2*, *3*, *4*, and *5*.

The equation describing differential equilibrium in roll torque about the operating point given in Fig *5*, is defined by equating the differentials of equations 8 and 9. Thus for stand *1*

$$\delta\tau_1=\delta\alpha_1+\beta_1\delta\omega_1=\frac{\partial\tau_1}{\partial h_0}\delta h_0+\frac{\partial\tau_1}{\partial h_{12}}\delta h_{12}+\frac{\partial\tau_1}{\partial f_{12}}\delta f_{12} \quad (16)$$

On rearrangement of terms to emphasize that δh_0 is a mill disturbance, equation 16 becomes

$$\delta\alpha_1-\frac{\partial\tau_1}{\partial h_0}\delta h_0=\frac{\partial\tau_1}{\partial h_{12}}\delta h_{12}+\frac{\partial\tau_1}{\partial f_{12}}\delta f_{12}-\beta_1\delta\omega_{12} \quad (17)$$

Roll torque equations similar to equation 17 are obtained for stands *2*, *3*, *4*, and *5*.

The five roll force equations in the form of equation 15 and the five roll torque equations in the form of equation 17 must satisfy the differential continuity equation 11. On the assumption that $V_3=v$, the peripheral velocity v of the rolls may be written in terms of the angular roll speed as

$$v=V_3=R\omega \quad (18)$$

where R is the undistorted roll radius. Equation 11 is rewritten in terms of angular roll velocity ω. For stands *1* and *2*, the differential continuity equation becomes

$$h_{12}R_1\delta\omega_{12}+V_{12}\delta h_{12}=h_{23}R_2\delta\omega_{23}+V_{12}\delta h_{23} \quad (19)$$

Three more independent continuity equations are obtained for stands *2* and *3*, *3* and *4*, *4* and *5*.

The incremental behavior of the mill is described by a total of 14 equations. In these equations there are 14 unknowns: five strip gauges, five strip speeds, and four strip tensions. This set of equations is given in matrix form in Table I. The term in any given row of the first column in this table is equal to the sum of the products of the strip variables with the coefficients in the same row. Thus for row 4,

$$-M_2\delta S_4=\frac{\partial F_4}{\partial h_{34}}\delta h_{34}+\left(\frac{\partial F_4}{\partial h_{45}}-M_4\right)\delta h_{45}+\frac{\partial F_4}{\partial f_{34}}\delta f_{34}+\frac{\partial F_4}{\partial f_{45}}\delta f_{45}$$

Mill Control Equations

The objective of this study has been to obtain equations expressing changes in the strip variables, $\delta h_{12}\ldots\delta h_{45}$, $\delta\omega_{12}\ldots\delta\omega_{23}$, $\delta f_{12}\ldots\delta f_{45}$ in terms of changes in the mill control variables, $\delta\alpha_1\ldots\delta\alpha_5$ and $\delta S_1\ldots\delta S_5$. These equations are obtained by solving the 14 equations given in Table I. The control equations have the form

$$\begin{aligned}
&b_{11}\delta S_1+\ldots b_{15}\delta S_5+b_{16}\delta\alpha_1+\ldots b_{1,10}\delta\alpha_5=\delta h_{12}\\
&b_{21}\delta S_1+\ldots b_{25}\delta S_5+b_{26}\delta\alpha_1+\ldots b_{2,10}\delta\alpha_5=\delta h_{23}\\
&\vdots\\
&b_{14,1}\delta S_1+\ldots b_{14,5}\delta S_5+b_{14,6}\delta\alpha_1+\ldots b_{14,10}\delta\alpha_5=\delta f_{45}
\end{aligned} \quad (20)$$

The matrix of b coefficients in equation 20 is the inverse of the matrix of coefficients in Table I.

The numerical evaluation of the b coefficients permits the quantitative evaluation of changes in strip variables caused by changes in control variables. For example, if δS_1 is altered by one unit while all other mill control settings remain the same, b_{11} is numerically equal to the change in gauge between stands *1* and *2*, b_{21} is numerically equal to the change in output gauge, etc. A different set of b coefficients is required to describe the incremental behavior of the mill for each distinctly different rolling schedule. In the next section, the numerical evaluation of the b coefficients is carried out for a typical rolling schedule on a representative steel mill.[15,16]

Numerical Evaluation of Incremental Mill Control Equations

Data for the mill schedule studied are listed in Appendix I.

Partial Derivatives for Roll Force F and Roll Torque τ

To calculate the roll force F and roll torque τ for a particular mill schedule, the following data must be known:

1. The mean yield stress of the material $\bar{k}$.
2. The radius of the work roll R.
3. The coefficient of friction μ.
4. The front and back tensions t_1 and t_2.
5. The input and output gauges h_1 and h_2.

The roll force F and roll torque τ were calculated from the nomograms of Ford and Lianis. To obtain the partial derivatives such as $\partial F_1/\partial h_{12}$ at the operating point, F_1 is determined for the operating value of h_{12} and for increments above and below this value. The results are plotted on a roll force F_1 versus thickness h_{12} diagram and a smooth curve is drawn through the calculated points. The slope

Table IV. 2-Stand Mill Control Equations

	δS_1, Thousandths of Inches	δS_2, Thousandths of Inches	$\delta\alpha_1$, Ton-Inch per Inch	$\delta\alpha_2$, Ton-Inch per Inch	δf_{12}, Ton per Square Inch
δh_{12}, thousandths of inches	0.738	−0.1140	0.00166	−0.0028	0.0522
δh_{23}, thousandths of inches	0.3697	−0.0289	0.0141	−0.024	−0.00611
$\delta\omega_{12}$, rpm	0.00218	0.02099	0.02342	0.00053	−0.00961
$\delta\omega_{23}$, rpm	0.00183	−0.05379	0.00067	0.07713	0.05986
δf_{12}, ton per square inch	−0.4738	1.5718	−0.02298	0.0391	−0.71928

Table V. Control Equations for a 5-Stand Mill

	δS_1, Thousandths of Inches	δS_2, Thousandths of Inches	δS_3, Thousandths of Inches	δS_4, Thousandths of Inches	δS_5, Thousandths of Inches	δm_1, Ton-Inch per Inch	δm_2, Ton-Inch per Inch	δm_3, Ton-Inch per Inch	δm_4, Ton-Inch per Inch	δm_5, Ton-Inch per Inch
δh_{12}*	0.737	−0.1138	0.0401	−0.0323	0.00368	0.00163	−0.00311	0.00187	−0.002025	0.000336
δh_{22}	0.3698	−0.029	−0.00469	0.00378	−0.00043	0.0141	−0.024	−0.000219	0.000237	−0.000039
δh_{32}	0.160	−0.0193	0.00855	−0.01398	0.0016	0.00622	−0.00608	−0.02337	−0.000877	0.000145
δh_{42}	0.152	−0.0184	0.00652	0.00455	−0.0120	0.0059	−0.000508	−0.000188	−0.0378	−0.001099
δh_{52}	0.1039	−0.0125	0.00443	−0.00348	0.01107	0.00403	−0.000345	0.0002	−0.000775	−0.03829
$\delta \omega_{12}$	0.00235	0.02094	−0.00738	0.00594	−0.00067	0.02343	0.000574	−0.00034	0.03727	−0.000062
$\delta \omega_{22}$†	0.000237	−0.0533	0.069	−0.0556	0.00633	0.000819	0.0767	0.00322	−0.03486	0.000578
$\delta \omega_{32}$	−0.0261	0.00152	−0.0252	0.1221	−0.0138	−0.000989	0.002461	0.3386	0.00765	−0.00127
$\delta \omega_{42}$	−0.04183	0.005017	−0.00248	−0.1952	0.253	−0.001022	0.000208	0.00978	0.7519	0.02311
$\delta \omega_{52}$	0.7068	−0.00856	0.00259	−0.0082	−0.4276	0.002742	−0.000332	0.00043	0.02194	1.5373
δt_{12}‡	−0.461	1.567	−0.532	0.455	−0.050	−0.0224	0.04295	−0.02583	0.0279	−0.00463
δt_{22}	−0.0177	0.0058	0.762	−0.6186	0.0705	−0.000742	−0.00522	0.0359	−0.0388	0.00544
δt_{32}	−0.152	0.0185	−0.00468	1.493	−0.170	−0.00589	0.000327	−0.0256	0.0936	−0.0155
δt_{42}	−0.290	0.0351	−0.0123	0.01487	1.5018	−0.01126	0.000962	−0.000824	−0.01807	0.1370

*$\delta h_{12} \ldots \delta h_{52}$ = thousandths of inches

†$\delta \omega_{12} \ldots \delta \omega_{52}$ = rpm

‡$\delta t_{12} \ldots \delta t_{42}$ = ton per square inch

of the tangent to this curve at the operating value defines the value of $\partial F_1/\partial h_{12}$.

Mill-Motor-Drive Damping Coefficients β

If the motor speed droop characteristic and the operating motor horsepower is known, β can be calculated. The calculation for β_2 on stand *2* is shown here.

The speed droop from no load to full load is taken to be 1.5 per cent and the motor is assumed to operate at 80 per cent of rated horsepower. From Appendix I, the schedule strip speed V_{22} is 600 feet per minute. The angular velocity of the rolls is given by

$$\omega = \frac{V_{22} \times 12}{2\pi R_2} = 112.0 \text{ rpm} \qquad (21)$$

where $R_2 = 10.25$ inches.

The full load torque is calculated presently

$$\tau_2 = \frac{(\text{hp})_2 \times 33{,}000 \times 12}{2\pi \times \omega_{22} \times 2{,}000 \times b} = 21.9 \text{ ton-inches per inch of strip width} \qquad (22)$$

$(\text{hp})_2$ = horsepower rating of mill motor, stand *2*, = 2,500 horsepower

$\omega_{22} = 112.0$ from equation 4

b = strip width = 32.75 inches

from Fig. 6 and from equation 21 it is estimated that the roll speed at full-load torque (point *B*) is 111.60 rpm and at no-load (point *C*) it is 113.3 rpm. Hence the damping coefficient is given by the following

$$\beta_2 = -\frac{BD}{DC} = \frac{-21.9 \times 60}{(113.3 - 111.6) \times 2\pi} = -121.22 \text{ ton-seconds} \qquad (23)$$

where BD is in ton-inch per inch of strip width, DC in radians per second. Values of β for the other stands are listed in Table II.

Estimation of Mill Stiffness Coefficient M

The mill stiffness M is required for calculations involving equation 6. The value of M has been estimated by the following procedure. The plastic roll force curve is drawn for stand *5*, for the mill schedule under study. The elastic line $F_5 = M_5(h_5 - S_5)$ must pass through the equilibrium point *A*. It is assumed that before the strip is threaded the roll surfaces are in contact so that $S_5 = 0$. The slope of the resulting elastic line drawn from point *A* to the origin in Fig. 7 is

$$M_5 = \frac{F_5}{h_5} = \frac{14.4}{0.0095} = 1{,}522 \text{ ton-inch per inch of strip width} \qquad (24)$$

Since all stands are identical, $M_5 = M_4 = M_3 = M_2 = M_1 = 1{,}522$ tons per square inch.

Evaluations of Constants in the Continuity Equations

The continuity equation for stands *1* and *2* is rewritten

$$h_{12}R_1\delta\omega_{12} + V_{12}\delta h_{12} - h_{22}R_2\delta\omega_{22} - V_{22}\delta h_{22} = 0 \qquad (25)$$

The values of h_{12}, h_{22}, etc., R_1, R_2, etc., and V_1, V_2, etc., are available from Appendix I. The incremental variables in equation 25 are $\delta\omega_{12}$, $\delta\omega_{22}$, δh_{12}, and δh_{22}. In Table II all constants for the four continuity equations are shown in suitable units. Thus linear velocities are in inches per second and the hR products in square inches.

Incremental Equilibrium Equations for the 5-Stand Mill

Table II represents the numerical evaluation of Table I, except for δh_0 which has been taken as zero. Inversion of the coefficient matrix in Table II will produce the desired control equations. It is of interest though, to examine the problem of obtaining these equations for a simple 2-stand mill.

Numerical Evaluation of Incremental Control Equation for a 2-Stand Mill

If the output strip from stand *2* of the 5-stand mill is assumed to be reeled at this point with controllable tension t_{22}, a 2-stand mill is obtained. The incremental control variables are δS_1, δS_2, $\delta\alpha_1$, $\delta\alpha_2$, and δt_{22}, while the strip variables are δh_{12}, δh_{22}, $\delta\omega_{12}$, $\delta\omega_{22}$, and δt_{12}. It is assumed that $\delta h_0 = 0$. The incremental equilibrium equations are shown in Table III, and the control equations corresponding to equation 20 are obtained by inversion of the matrix in Table III. The inversion was carried out by hand computation using the elimination method outlined by Hartree.[17]

Numerical Evaluation of Incremental Control Equations for a 5-Stand Mill

The required equations were obtained by inversion of the 14×14 matrix in Table II. The digital computer at the University of Toronto was used to carry out the necessary computations.

Interpretation and Application of Mill Control Equations

The 2-Stand Mill

From Table IV the change in output gauge for the 2-stand mill, namely δh_{22}, is given as

$$\delta h_{22} = 0.3697\delta S_1 - 0.0289\delta S_2 + 0.0141\delta\alpha_1 - 0.024\delta\omega_2 - 0.0061\delta t_{22} \qquad (26)$$

It should be noted that the coefficient of δS_2 in equation 26 is negative. Thus an

increase in the separation of the rolls in stand 2, causes a decrease in the output gauge. This apparent anomaly is explained by the fact that as δS_2 is increased, the interstand tension t_{12} increases by the increment $\delta t_{12} = 1.57\delta S_2$ and the gauge coming out of stand 1 falls by $\delta h_{12} = 0.11\delta S_2$. While a unit increase in δS_2 causes a notable increase in δt_{12}, a unit increase in δS_1 causes a decrease in interstand tension δt_{12}.

This observation suggests a method of output gauge control, that will maintain t_{12} constant. If $\delta\alpha_1$, $\delta\alpha_2$, and δt_{12} in Table IV are taken at zero, the control equations in terms of δS_1 and δS_2 become

$$\begin{aligned}
\delta h_{12} &= 0.74\delta S_1 - 0.11\delta S_2 \\
\delta h_{23} &= 0.37\delta S_1 - 0.03\delta S_2 \\
\delta\omega_{12} &= 0.002\delta S_1 + 0.02\delta S_2 \\
\delta\omega_{23} &= 0.002\delta S_1 - 0.05\delta S_2 \\
\delta t_{12} &= -0.47\delta S_1 + 1.57\delta S_2
\end{aligned} \tag{27}$$

From equation 27 it follows that

$$\delta t_{12} = 0 \text{ when } \delta S_1 = 3.34\delta S_2 \tag{28}$$

If the condition of equation 28 is introduced in equation 27 the control equations in terms of δS_1 become

$$\begin{aligned}
\delta h_{12} &= 0.707\delta S_1 \\
\delta h_{23} &= 0.361\delta S_1 \\
\delta\omega_{12} &= 0.008\delta S_1 \\
\delta\omega_{23} &= -0.005\delta S_1 \\
\delta t_{12} &= 0
\end{aligned} \tag{29}$$

Equation 29 shows that a method of gauge control preserving constant interstand tension is realizable if the roll settings of the two stands are altered simultaneously. This result is typical of the insight into mill operation that the control equations yield.

The 5-Stand Mill

The results of Table V yield a detailed picture of the steady-state incremental behavior of the mill for the particuar rolling schedule examined. While it has not been possible to check experimentally the accuracy of the coefficients, the values given are in general accord with mill operating experience. Major significance is attached to the relative magnitudes of various control coefficients. For example, Table V shows that adjustment of output gauge h_4 with δS_4 or δS_5 will probably cause strip breakage, because the per-unit changes in tension δt_{34} and δt_{45} associated with δS_4 and δS_5, respectively, are rather large compared to the corresponding per-unit changes in δh_4. A similar comment applies to the use of δS_3.

From a study of the control coefficients for stands 1 and 2, it is apparent that simultaneous adjustment of δS_1 and δS_2 in a suitable ratio can lead to a sensitive means of control of output gauge that maintains constant tension between stands 1 and 2. This method was indicated for the 2-stand mill, and can be applied to the 5-stand mill. Examination of control coefficients associated with stalled-torque variations $\delta\alpha$ in the mill motors indicates that only $\delta\alpha_3$ has any significant effect on output gauge. This fact appears to have particular significance for the development of precise automatic gauge control.

Table VI. Mill and Schedule Data

	Stands				
	1	2	3	4	5
Work roll diameter, inches	20.0	20.5	19.6	21.0	21.0
Work roll length, inches	56.0	56.0	56.0	56.0	56.0
Mill motor rating, horsepower	1,500	2,500	2,500	2,500	3,000
Mill motor speed range, rpm	70–175	125–250	175–400	225–525	250–625
Strip speed range, feet per minute	385–965	688–1,376	965–2,200	1,238–2,880	1,376–3,400
Input gauge, inches	0.080	0.0662	0.0331	0.0191	0.0134
Output gauge, inches	0.0662	0.0331	0.191	0.0134	0.0095
Draft, inches	0.0138	0.0331	0.0140	0.0057	0.0039
Reduction, per cent	17.25	50.0	42.3	29.8	29.1
Front tension,* ton per square inch	13.0	16.8	13.1	13.4	14.0
Width of strip, inches	32.75	32.75	32.75	32.75	32.75

*All ton units are in long tons to conform with the units used in the nomograms.

Table VII. Operating Point Data for Schedule of Table VI

	Stands				
	1	2	3	4	5
Output strip velocity, feet per minute	300	600	1,040	1,480	2,100
Roll speed, rpm	57.4	112.0	202.3	269.4	382.0
Motor damping coefficient, ton-seconds	−402.44	−121.22	−27.72	−12.335	−5.984
No-load motor speed, rpm	57.8	113.3	205.8	273.6	390.0
Stalled motor torque,* ton-inch per inch	2,540	1,460	605	364	244
Coefficient of friction μ	0.05	0.05	0.05	0.03	0.025
Mean yield stress,* ton per square inch	32.3	46.2	48.5	50.4	51.0

*All ton units are in long tons to conform with the units used in the nomograms.

Conclusions

Steady-state incremental control equations have been derived and numerically evaluated for a 5-stand cold-rolling steel strip mill. These equations are obtained from the theory of plastic deformation in rolling and the electromechanical equilibrium equations for the mill. It has been shown that the numerical evaluation of the equations is practicable. The numerical form of the equations suggests preferred forms for automatic mill control systems and provides an analytical basis for studying the performance of such systems. For example, Table V suggests that δS_1, δS_2, and $\delta\alpha_3$ are the most suitable control variables in the mill studied. Because of assumptions made in the course of the numerical evaluation, it is not known how accurately the control coefficients represent mill behavior for the schedule studied. The theoretical basis for the equations is considered to be adequate for control studies. In the numerical examples given, importance is attached to the relative magnitudes of the coefficients. These appear to be largely in accord with manual mill operating experience. The study indicates that there is a minimum level of mill instrumentation required if the dynamic performance of the mill is to be clearly understood.

The study illustrates the nature of the multivariable control problem that is of increasing importance in system design.

Appendix

Table VI gives some general data on a 5-stand tandem cold mill including the rolling schedule for which the analysis was carried out.[15,16] Table VII contains the operating values for mill variables calculated for the schedule in Table VI.

Nomogram Calculations for Roll Force and Roll Torque

Roll forces and roll torques and their derivatives were calculated for the schedule of Table VI with assumptions listed in Table VII, using the nomograms of Ford and Lianis. The values of mean yield stress in Table VII were calculated from the curve of yield stress versus reduction supplied with the nomograms. The partial derivatives such as $\partial F_i/\partial h_{ji}$ were approximated from nomogram evaluation of F_i at h_{ji}, $h_{ji} \pm \Delta h$.

To start the computation it is necessary to assume a value for the deformed roll radius R'. The calculations lead to a value

for the roll force F and a distorted roll radius R_1'. If R_1' is not equal to R' assumed, the calculation must be repeated until both assumed and calculated values are equal. Ford[18] has outlined a procedure for the finding of the roll force F which requires only two trial calculations. The value of R' determined for the roll force calculation is used in the determination of the roll torque.

References

1. Network Analysis and Feedback Amplifier Design (book), **H. W. Bode.** D. Van Nostrand Company, Inc., New York, N. Y., 1945.

2. Theory of Servomechanisms (book), **H. M. James, R. S. Phillips, N. B. Nichols.** *MIT Radiation Laboratory Series, No. 25*, McGraw-Hill Book Company, Inc., New York, N. Y., 1947.

3. Principles of Servomechanisms (book), **G. S. Brown, D. P. Campbell.** John Wiley & Sons, Inc., New York, N. Y., 1948.

4. Automatic and Manual Control, **A. Tustin.** Butterworth Scientific Publications, London, England, 1952.

5. Control System Synthesis (book), **J. Truxal.** McGraw-Hill Book Company, Inc., New York, N. Y., 1955.

6. General Algebraic Method Applied to Control Analysis of Complex Engine Types, **A. S. Bokseabom, R. Hood.** *Report 980*, National Advisory Committee on Aeronautics, Washington, D. C., 1950.

7. Mathematical Theory of Multi-Dimensional Servo Systems, **M. Colomb, E. Usdine.** *Journal*, Franklin Institute, Philadelphia, Pa., vol. 253, Jan. 1952.

8. A Method for the Preliminary Synthesis of a Complex Multi-Loop Control System, **D. J. Povejsil, A. M. Fuchs.** *AIEE Transactions*, vol. 74, pt. II, July 1955, pp. 129–34.

9. Principles of Continuous Gauge Control in Sheet and Strip Rolling, **W. C. F. Hessenberg, R. B. Sims.** *Proceedings*, Institution of Mechanical Engineers, London, England, vol. 166, 1952, p. 75.

10. The Calculation of Roll Pressure in Hot and Cold Flat Rolling, **E. Orowan.** *Ibid.*, vol. 150, 1943, p. 140.

11. The Calculation of Roll Force and Torque in Cold Strip Rolling with Tensions, **H. Ford, D. R. Bland.** *Ibid.*, vol. 159, 1948, p. 144.

12. Cold Rolling with Strip Tensions, **H. Ford, F. Ellis, D. R. Bland.** *Journal*, Iron and Steel Institute, London, England, vol. 168, 1951, p. 57; Vol. 171, 1952, p. 239.

13. Beitrag zur Theorie des Walzvorganges, **Th. V. Karman.** *Zeitschrift für Angewandte Mathematic und Mechanic*, Berlin, Germany, vol. 5, 1925, p. 139–141.

14. Incremental Control Equations for Multiple Stand Mills, **J. H. Courcoulas.** *M.A.Sc., Thesis*, Department of Electrical Engineering, University of Toronto, Toronto, Ont., Canada, 1955.

15. Electrical Equipment for Cold Rolling and Finishing Department, **M. A. Leishmann.** *Iron and Steel Engineer*, Pittsburgh, Pa., vol. 27, Aug. 1950, p. 45.

16. Electrical Drive for 56 In. Five-Stand Tandem Cold Reduction Mill, **R. A. Harvie.** *Ibid.*, p. 49.

17. Numerical Analysis (book), **D. R. Hartree.** Oxford Clarendon Press, Oxford, England, 1952.

18. Researches Into the Deformation of Metals by Cold Rolling, **H. Ford.** *Proceedings*, Institution of Mechanical Engineers, London, England, vol. 159, 1948, p. 115.

Discussion

Discussion of Paper 56-638

R. M. Saunders (University of California, Berkeley, Calif.): From the control system viewpoint there are several omissions or unstated assumptions omitted from Paper 56-638; clarification of the following will add materially to its completeness.

First, the paper seemingly ignores the dynamics of one of the two degrees of freedom present in the mechanical coordinates of each of the stands. Mention is made of the dynamics of the mill housing which covers one of the degrees of freedom, the vertical motion of the housing. However, no justification is made for omitting the moment of inertia of the drive motor and the rolls. While it may be true that the speed of the stand drive motor is regulated in a similar fashion to the reel motor, some statement should be made as to the resultant corner frequencies, time constants, or pole-zero locations to justify their omission from the analysis.

Second, in using incremental techniques the author is using a terminated Taylor series expansion about a point. Since the complete Taylor series involves second- and higher-order derivatives, justification for neglecting second- and higher-order terms must lie in magnitude of the perturbation. Some statement should be made giving the bounds between which the analysis is valid.

Finally, no experimental results are presented in spite of the many tacit and implied assumptions made in the analysis. Computer studies are essential in applying controls to a complex process such as a steel mill, but are not the end in themselves. Whether the analyses and assumptions lead to systems which met all the specifications is the real issue in point.

James M. Ham (University of Toronto, Toronto, Ont., Canada): Mr. Phillips has derived equations that describe the transient incremental equilibrium of tandem rolling mills. He shows a computer simulation of the equations for the last two stands of a 5-stand mill and uses this to determine the optimum adjustment of a particular scheme for vernier gauge control.

The author is to be commended for establishing a fundamentally sound basis for his control studies by applying rolling theory. His paper is an excellent illustration of how deeply electrical control engineers must be concerned with the characteristics of the loads their motors drive if they are to have a rational basis for control system design. To my knowledge, his paper and our own paper[1] represent the first published numerical studies of this kind on rolling mills. Hessenberg and Jenkins[2] previously had considered some aspects of the incremental behavior of tandem mills in a qualitative manner. These studies have been greatly facilitated by the outstanding work of Ford, Bland, and others in England in developing rolling equations suited to practical applications.

Mr. Phillips has been concerned with adjusting the performance of a particular control scheme and therefore has not used general mill control equations that relate increments in every mill control variable (screws and motor adjustment on every stand) to mill response variables (strip gauge, tension, and speed in every stand). These equations are found by taking the inverse of the matrix of the equilibrium equations written for all stands simultaneously. The general form of these equations for steady-state conditions is given in our paper with a numerical example. These general equations permit various possible control schemes to be considered, and in particular cast some interesting light on Mr. Phillips' study.

The control scheme described by Mr. Phillips apparently is intended to act essentially as two regulators in cascade. For purposes of adjusting the fine regulator operating on stand 5, the assumption is made that operation of the fine regulator does not interact significantly with the early stands in the mill. A typical quantitative justification for this assumption is obtained from Table V of our paper[1] by considering the relative magnitudes of mill control coefficients associated with the motor adjustment at stand 5, namely α_5. However, it is important to note that operation of the coarse regulator loop on stand 1 propagates effects, in particular important changes in strip tension, throughout the mill. Feedback analysis based on our general mill equations reveals the possibility of important tension changes when the gauge of the input strip changes, even when the coarse regulator loop is holding the gauge out of stand 1 essentially constant.

Since tensions as well as gauges are important in a tandem mill, the foregoing considerations beg the question as to whether there are mill control schemes, different from that described by Mr. Phillips, that are capable of controlling tensions as well as gauges within suitable tolerance bounds. Multivariable control theory applied to the general equations developed in our paper shows that there are many such schemes possible. One such scheme is shown in Fig. 1 of this discussion. The distinctive features of this system are that use is made of (1) values of the input strip gauge measured before the strip enters the mill, (2) more than one screw adjustment, (3) auxiliary variables such as roll force,[3,4] roll torque, strip tensions. The function of the computer is to weight the computer inputs in such a way as to realize the over-all gauge-tension control desired. Study of the practical feasibility and economic worth of such schemes deserves attention. It is not likely that the best gauge control in tandem mills can be obtained by using gauge information alone. For example, the use of roll force measurements which has been pioneered and patented by the British Iron and Steel Research Association largely eliminates the problem of transportation lag associated with exit gauge measurements. Mr. Phillips' scheme uses gauge information alone and although it has seen actual service

it is not clear to what extent his system achieves gauge control in practical operation.

In his study, Mr. Phillips deals with the incremental behavior of a control scheme operating about set points. The major loss through off-gauge tonnage in modern mills continues to occur during the acceleration and deceleration periods associated with the beginning and end of the coil. When the mill speed is changed there is the well-known "speed effect" that causes mill control coefficients to change and hence gauges and tensions to change. Further, there is often transient change in input strip thickness, between the end of one strip and the beginning of the next welded to it.

It is not clear that the control scheme presented by Mr. Phillips, or for that matter any of the many possible incremental control schemes, can operate effectively from threading to full speed and down again. I would appreciate his discussion of the performance of his control scheme during these periods.

REFERENCES

1. INCREMENTAL CONTROL EQUATIONS FOR TANDEM ROLLING MILLS, J. H. Courcoulas, J. M. Ham. *AIEE Transactions*, vol. 75, pt. II, 1956 (Jan. 1957 section), pp. 363–74.

2. EFFECTS OF SCREW AND SPEED-SETTING CHANGES ON GAUGE SPEED AND TENSION IN TANDEM MILLS, W. C. F. Hessenberg, W. N. Jenkins. *Proceedings*, Institution of Mechanical Engineers, London, England, 1955–56.

3. IMPROVEMENTS IN AND RELATING TO THE PRODUCTION OF METAL AND OTHER SHEET AND STRIP, W. C. F. Hessenberg, R. B. Sims, *British Patent no. 692,287*, June 3, 1953.

4. *British Patent no. 713,105*.

Discussion of Paper 56-656

R. M. Saunders (University of California, Berkeley, Calif.): Paper 56-656 shows every evidence of having been carefully constructed and illustrates how the systematic analysis of control system technology will solve complex problems in processes. While some may criticize the authors for the use of certain rolling mill equations or question the validity of their assumptions, the methods developed by the authors are sound and will yield results for any set of assumptions or mill equations.

There are a couple of points that should be clarified. While the authors start their analysis for the dynamic case with equation 1, their results are actually for the steady-state case for $s = 0$ in the Laplace equations. Thus, they develop only the special case of the transfer functions H_F and H_T* where all derivatives with respect to time are negligibly small. Second, they do not adequately warn the reader that he must check the regions of validity rather carefully. In using incremental techniques, the authors are using a terminated Taylor series expansion about a fixed operating point. Since the complete Taylor series involves second- and higher-order derivatives, justification for neglecting these terms lies in the magnitude of the perturbation. Thus a test should be made to see if all excursions about fixed operating points really justify a single derivative termination of the series.

One of the most significant aspects of this paper is the contribution to control system theory made by the methods described by the authors. The problem of including the mutual effects between co-ordinate systems is difficult indeed. The authors' use of incremental equations and inversion of the matrix should point the way to many others wishing to solve similar problems.

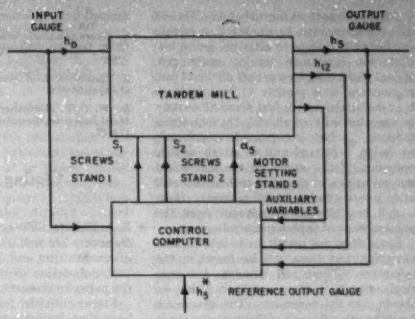


Fig. 1. Multivariable mill control system

H. S. Maxwell (General Electric Company, Schenectady, N. Y.): The authors conclude from Table V that the roll separation of stand *1*, the roll separation of stand *2*, and the torque of stand *5* motor are likely to be the most suitable control variables for a 5-stand tandem mill. Presumably it is intended that all three be used.

We have two regulating systems in operation in the Pittsburgh, Pa., area and are building many more. Such equipment is popularly known as automatic gauge control. Thus far we have chosen to use two of the three control variables suggested by the authors.

Roll separation of stand *1* is used as one control variable. An X-ray thickness gauge measures the thickness of strip leaving stand *1* and operates the d-c screwdown control of stand *1* to regulate thickness leaving stand *1*. This part of the total system is known as the coarse regulator.

Torque of stand *5* motor is used as a second control variable. A second X-ray thickness gauge measures the thickness of strip leaving stand *5*. This part of the system is known as the vernier regulator.

The possibility of using the roll separation of stand *2* as a third control variable is interesting. Before doing so, however, as an electrical manufacturer we would consider it necessary to make a transient analysis of the effects of this control variable on interstand tensions. For example, transient values of tension may appear between stands *1* and *2* or between stands *2* and *3* which would cause an adverse effect on thickness over a small part of total length. This might offset to some extent the advantage of using this control variable. Such questions can only be resolved by transient analysis, using an analogue computer.

Perhaps of interest with regard to the use of roll separation as a control variable is the need to return to a starting value of roll separation after each coil of an order, to be ready for the next coil. This function is provided by a precise position regulator interlocked to occur automatically.

Such equipment is provided for the stand *1* coarse regulator and if the roll separation of stand *2* were also used as a control variable, a similar screw return regulator would be needed. The authors have had no occasion to study economic considerations, but as a matter of interest the addition of the third control variable would probably increase the cost of present automatic gauge control about 50 per cent.

Discussion of Papers 56-638 and 56-656

W. N. Jenkins and **W. C. F. Hessenberg** (The British Iron and Steel Research Association, London, England): We were particularly interested in the two papers on tandem mills. We have ourselves covered some of the ground described in the Ham and Courcoulas' paper, 56-656, including the derivation of the equilibrium equations for a 5-stand mill,[1–3] and have since had the privilege of personal discussions with Mr. Courcoulas.

Our approach, however, has been somewhat different from his. We have been less concerned with the application of control systems to existing tandem mills than with the prospect of providing a rational basis for improved mechanical and electrical design of new mills. We have also been conscious of the fact that the methods of roll force and torque calculation available to us, including the Ford and Lianis nomograms, do not deal adequately with the extreme conditions of rolling found on the fourth and fifth stands of a tin-plate mill and we should be interested to know how the authors surmounted this obstacle.

We have therefore concentrated mainly on a qualitative approach and have found that it can yield a good deal of information. For example, the authors have pointed out that lifting the screws of the second stand of a 2-stand mill can cause a decrease in output gauge. Our analysis not only shows this, but also the fact that if the damping coefficient of the drive (motor droop) is sufficiently increased, the opposite effect is obtained, i.e., a lifting of the screws increases the gauge. It also reveals that it is only on even-numbered tandem mills that this "anomalous" effect on the last stand occurs; with 3- and 5-stand mills, raising the screws of the last stand invariably increases the gauge.

Again, the authors show that on a 2-stand mill the screws can be worked in a certain combination such as to alter the gauge but leave the interstand tension unchanged. We may add that, by a proper choice of mill parameters, it is possible to arrange it so that the screws on the first stand alone may be operated without altering the interstand tension. In fact, it is possible in principle to design a 5-stand mill in which neither operation of the screws on the first stand nor variations of the wind reel tension will alter any of the interstand tensions. Such a mill has obvious advantages from the point of view of applying control systems.

Space does not permit us to add further examples, but these will be found in the references. There are, however, two particular points in the paper upon which we should like to comment. The procedure for estimating the mill stiffness coefficient is decidedly unsound and leads to a figure which is too small by a factor of at least ten for this type of mill. The error arises from the assumption that the elastic line is straight all the way down to zero load. In fact it usually has a pronounced tail such that the slope is very small at low loads becoming nearly linear and very much steeper than shown in the authors' Fig. 7 in the working range.

The other point concerns the forward slip which the authors dismiss with the approximation of their equation 18. Some tentative calculations that we have made for a 3-stand mill suggest that the contribution of forward slip to the changes in strip speed are by no means negligible.

In regard to Paper 56-638, we would like to congratulate Mr. Phillips on his resolute attack on the problem of transient changes in tandem mills. We had thought a little about this ourselves but not very seriously, having decided that just as it is desirable to know something about the implications of Ohm's law before tackling the problem of transient changes in electric circuits, so the equilibrium states of a tandem mill should be understood before its transient behavior is approached. Mr. Phillips has taken a short cut and, within the bounds of his particular problem of applying an automatic gauge control to an existing mill, he is probably justified.

With reference to the second of Mr. Phillips' conclusions, we are doubtful about the assumption that neither the speed nor the thickness out of stand *4* is affected by the vernier control on stand *5*. Our view is that the volume flow of metal through stand *4* is what remains practically unchanged and as a consequence the fractional-speed change of this stand will be equal and opposite to the fractional gauge changes out of it. In fact, we should like to emphasize that the significant dependent variable in tandem mill dynamics is the volume flow rate of metal, or "output" as we have called it, and the changes in speed and gauge are derived from it.

We note that Mr. Phillips, too, has been using the customary methods of roll force and torque calculation and applying them to the last two stands of a tin-plate mill. This procedure by itself is difficult to justify and we presume that he has supplemented his calculations with measurements made during the extensive testing of the mill itself of which he writes in his opening paragraph.

REFERENCES

1. Correspondence from **W. C. F. Hessenberg, W. N. Jenkins.** *Research*, London, England, Aug. 1954.

2. Reference 2 of discussion by **J. M. Ham** of Paper 56-638.

3. **W. C. F. Hessenberg, W. N. Jenkins.** *Sheet Metal Industries*, London, England, Oct. 1955, pp. 771–77.

Closing Remarks

R. A. Phillips: The comments of the various discussers are well chosen. This is a complicated subject and one which requires further elaboration of certain points made in the paper to answer the questions raised.

I agree with Mr. Jenkins and Mr. Hessenberg that further analytical work on the rolling process would be desirable. The investigations of these two men and their colleagues in the British Iron & Steel Research Association which have been reported in numerous papers have been extremely useful. The publications have indicated that the theories developed are well substantiated by test. A strong need for transient analysis was felt 2 years ago to study theoretically problems in automatic gauge control and the best available theories of single-stand cold-reduction mill behavior were used. When further work indicates useful refinements in the theories, they will certainly be included.

The question of whether the second conclusion of the paper is justified can be resolved for a particular mill by following the recommendation of conclusion 3. The point about volume flow of metal is an interesting one. If the steady-state behavior of a tandem mill is analyzed using the steady-state form of the equations in the paper, it can be shown that if the stand *5* motor is speeded up 1 per cent and the other stand motors do not change speed and the winding reel holds constant tension, the gauge out of stand *5* is decreased 0.98 per cent. The gauge out of stands *1*, *2*, *3*, and *4* change $+0.004$, -0.013, $+0.049$, and -0.121 per cent respectively. This indicates that the principal feedback loop in the mill for a stand *5* vernier control is one which could be deduced from the "mass flow" concept. That is, that a 1-per-cent speed change of stand *5* motor produces a -1-per-cent thickness change out of stand *5*. Transiently, however, the secondary effects can be important. For instance, the winding reel must change speed again to reach the reference tension (or current); the -0.212-per cent change in thickness out of stand *4* does not appear immediately but must build up as the strip propagates from stand to stand; the stand *4* load changes as the tension changes which will cause its speed to vary. However, since these effects are secondary in nature and the regulating system is self-correcting (being a feedback control system), it is not necessary to know the constants of the subsidiary loops to a high degree of accuracy. Therefore, I used what I thought, and still think, to be the best available calculation procedure to evaluate these constants. Nothing has arisen in the extensive field testing that could not be explained with the transient equations presented here.

I assume that the criticism by Mr. Jenkins and Mr. Hessenberg of the method of estimating mill stiffness applies only to the Courceulas-Ham paper.

In reply to Professor Saunders' questions, it was stated in the paper that it is necessary to consider the transient characteristics of the drive motors. The moment of inertia of the drive motor and rolls must be considered as must the motor-field time constant and the armature inductive time constant and the parameters were included in all studies. The system of equations describing the transient behavior of a d-c motor are well defined in the literature and it was felt that a repetition here would be of questionable value.

The use of first-order incremental techniques must, of course, be justified on a numerical basis. It can be expected that the variations in thickness of a well-regulated mill will be well within 5 per cent and the first-order approximation is justified for those coefficients which relate to thickness changes. The tension changes required to regulate gauge vary over a substantial range in some class but fortunately the coefficients which relate to tension changes are nearly linear from zero tension to a point where the strip would fracture between stands. As was pointed out in answer to another question, the use to which the equations are put justifies much of what was done.

The analytical insight into the rolling process and the graphical results of the computer studies have been very valuable in making successful applications. The equipments which have been installed have performed as expected. Reference thickness changes have been introduced and the resulting response has checked computer results very well. Where there has been difficulty, the knowledge of how the system should perform has made it possible to find the trouble spots much more rapidly. Field testing and adjustment is done on the best computer, the mill itself; but it is the most expensive one unfortunately and should be used sparingly. The tests which were performed on the system described in this paper were reported in reference 1 of the paper.

Professor Ham and his colleague, Mr. Courceulas, have chosen to study different aspects of the common problem of theoretical behavior of tandem cold reduction mills and have found that a steady-state analysis is sufficient for their purposes. The "general mill equations that relate increments in every mill control variable to mill response variables" are a special case of the general transient equations and as such are included in my paper. Take the numerical values for each stand as given in Table I and substitute them into equations 21, 22, and 48. If it is recognized that the delivery thickness and tension of one stand are the entry thickness and tension of the next stand in these equations and in equations 38 through 42 a matrix like the one described in Professor Ham's paper results. I have also explored alternate methods of gauge control using matrix inversion techniques. What Professor Ham states about the tension variations produced in the mill due to a change in screw setting of the first stand screws is true as far as it goes. In the paper to which he refers, the variation of entry gauge to the first stand was neglected. Without this variation there would be no need to move the stand *1* screws. From the inverted matrix including entry gauge

variation, I have found that if there is no variation of mill motor speed with load, a 5-per cent increase in entry gauge produces a decrease in interstand tension throughout the mill that is in the order of 2,000 pounds per square inch. If the thickness change out of stand *1* (4.0 per cent) is eliminated by a 7.5-mill lowering of the stand *1* screws, the interstand tension will be brought back nearly to its original value. Thus, the "coarse" control substantially reduces the tension variations which would be present without it. The usual practice on modern high-speed tandem mills is to supply each mill motor from its own generator. The terminal voltage of each of the generators is individually regulated. Provision is made to compensate for motor current-resistance drop a successively greater amount in each succeeding stand. The consequence is that as the entry gauge increases, the load on each of the mill stands increases. The tapered speed changes in the various stands tend to maintain interstand tensions and thus act to maintain gauge inherently. There is no field experience to indicate that the stand *1* coarse control is generating disturbances as Professor Ham suggests. Where entry variations of strip thickness are substantial, the coarse control reduces the delivery gauge variation considerably.

The transportation lag of the stand *5* X-ray gauge is not the limiting item in the vernier gauge control speed of response. To use the British Iron & Steel Research Association roll-force method of calculating delivered thickness for the vernier control does not seem justified. Strip tolerances can be as low as ±1 per cent. Therefore, it is difficult to see how the roll-force method can be used without an automatic calibration of it against a more absolute reference such as an X-ray gauge.

The best system in the eyes of the steel plant operator is not necessarily the optimum one from the theoretical or analytical standpoint, but the least complicated one that achieves practical satisfactory results. Addition of an incoming X-ray gauge, movement of several sets of screwsdowns, and use of roll pressure devices would conceivably improve the theoretical accuracy of the gauge control system. However, the accuracy of the system described in the paper is limited by the accuracy of available components and measuring methods rather than system design.

The problem of controlling gauge during acceleration is one on which my company is presently working. The automatic gauge control as described here brings the delivered strip gauge on tolerance at about 1,600 feet per minute on a mill rated at 3,850 feet per minute which is a substantial improvement over the result heretofore (see reference 1 of the paper.) The only reason that the gauge control does not bring the strip on gauge at a lower speed is that the tension between stands *4* and *5* would exceed the limits of safe rolling practice. Clearly, the "speed effect" and the problem of the large gauge variations at the welds are the next gauge control problems to be solved. They will not necessarily be solved with a feedback control system.

J. H. Courcoulas and J. M. Ham: We are grateful to Professor Saunders for emphasizing the fact that the matrix inversion technique we used for determining control equations has importance in itself, quite apart from the particular application to steel mills. This technique has been described in some detail.[1] If control engineering for multivariable machines and processes is to progress substantially beyond the now common use of cascaded setpoint regulators which often interact with one another, the multivariable control characteristics of the process or machine must be placed clearly in evidence. The matrix inversion technique applied to incremental equilibrium equations makes possible detailed control analysis of elaborate multivariable systems of which the modern tandem steel mill is an excellent example.

Professor Saunders quite correctly points out that whereas we professed to be finding general control transfer functions, we actually evaluated only the $S = 0$ or steady-state zero frequency values of these functions. In multivariable control problems a major step is of course to design for satisfactory steady-state behavior in response to step changes in variables. This step can be carried out with the data given. We have been preparing a second paper dealing with the general character of the transfer functions. Mr. Phillips in Paper 56-638 has provided a treatment of the transient problem in terms of equilibrium equations, but has not dealt with the inverse matrix of control equations. Since the transfer functions for the equilibrium equations involve mainly pure time lags, the general matrix inverse can be fairly readily obtained. It reveals, by the way, the characteristic wave-like propagation of gauge and tension changes through the mill.

To the question of the validity of incremental analysis which represents functions of several variables by terminated Taylor series, we cannot give a satisfying answer. Its validity depends very much on the particular application. The great practical merit of the method is that it renders highly complex systems amenable to analysis in a limited form. In selecting values for derivatives we used mean values over a ±10-per-cent change in the independent variable. It is interesting to observe that the change in many, but not all, of the rolling variables was well represented by such a mean value. A ±10-per-cent range is satisfactory for normal mill control conditions which are inherently incremental in nature. The acceleration and deceleration periods are not adequately described by invariant incremental equations and constitute the outstanding challenge of mill analysis.

We appreciate having the comments of Mr. Hessenberg and Mr. Jenkins. It is particularly appropriate that their discussion should record references to their own studies. We have been concerned primarily to show how to set up general incremental control equations that can provide a rational basis for the design of mill control systems, old and new. The development of general control equations from equilibrium equations by matrix inversion had not been published. The equations given by Mr. Hessenberg and Mr. Jenkins in reference 2 of their discussion are not adequate for control system design. They state that their concern has been to provide a rational basis for improved mechanical and electrical design. Since mill controllability is of fundamental importance, there appears to be no better way in which to evaluate the effects of varying such parameters as mill stiffness and motor droop than in terms of a general set of control coefficients as provided by our method of analysis. We have done some study in this direction.

The discussers point out that several aspects of the particular quantitative example given in our paper agree with the qualitative results they have given. They go on to comment on certain other general characteristics possible in mills. In this regard, their studies contribute much to an understanding of the forms of behavior that may be expected of tandem mills. Interestingly enough, the matrix inversion technique used in our paper can be reversed. That is to say, a set of desired control coefficients can be selected and by inversion of the control matrix, the values of the mill design parameters necessary to achieve the defined control characters can be deduced. Such a general synthesis of a mill design would require a great deal of knowledge as to what it is reasonable to expect from a mill. The discussers are far better equipped than ourselves to exploit this technique.

Mr. Hessenberg and Mr. Jenkins raise a number of questions regarding our application of rolling theory. They point out that we neglected forward slip. We appreciate that it may well be of importance. It is a simple matter to consider this factor as far as the general matrix method of analysis is concerned. Ford has recently included provision for slip, and its partial derivatives with respect to strip parameters, in a manuscript on the determination of partial derivatives of rolling variables. At the time our paper was prepared, we had no rolling analysis suitable for taking adequate account of slip.

On the question of the applicability of rolling theory as presented by Bland and Ford to the fourth and fifth stand of a 5-stand mill, we are not competent to judge. The hardness of the strip in these stands and the relatively large src of elastic recovery is such that this theory may not be applicable in an unmodified form. The distinguishing feature of our use of this theory is that we have used a coefficient of friction much lower than is usual. The major error arising would appear to be in torque calculations.

On the specific question of the technique we used to estimate mill stiffness, the following comments are in order. We appreciate that the curve of load force versus screw setting has a pronounced curved tail at low values of roll force, and, had we estimated mill stiffness by our technique in, say stand *1*, the value obtained would have been low by a factor of perhaps ten. However, the estimation was based on stand *5* where the strip is quite hard and the intercept on the gauge axis of the straight line that extends the slope of the elastic line at the operating point may in fact be negative. Hence our value for mill stiffness is likely high rather than ten times low. Calculation shows that the housing stiffness alone on the mill we examined is about five times our figure for mill stiffness. Since the roll stiffness is in parallel with the housing stiffness and is much lower than the housing stiffness, the mill stiffness is on the order of our figure, probably lower.

Mr. Hessenberg and Mr. Jenkins quite properly have questioned specific facets of

our use of available rolling theory. We would conclude by observing that the general method of determining control equations by the inversion of the matrix of equilibrium equations is sound and can be used with any rolling theory that includes, if desired, the effects of forward slip. Until a complete rolling theory is developed, we are inclined to use what is available to derive quantitative results that may be quite approximate. In designing mill control systems, even approximate control coefficients are of immense value. What appears to be most needed at present is more adequate mill instrumentation on representative tandem mills. With such instrumentation analytical results can be put to the acid test of experiment.

Mr. Maxwell raises the question of our suggesting that a particular group of three control variables is the most suitable for a 5-stand mill. For the numerical example given, these variables represent a possible choice particularly if the control of interstand tensions as well as gauges is to be considered. From a control standpoint we would emphasize our conviction that optimum control of tandem mills will require the adjustment of all controlling variables in a related manner. Hence, in our opinion, the cascading of simple regulators which actuate a single control variable at a time will not lead to the best gauge control system. For this reason, Mr. Maxwell's point that the use of a third control variable would add to the cost of a gauge control system is true but not exactly relevant. We have studied economic conditions in steel production enough to know that immense returns are to be had by improving on-gauge yield. The economic worth of the many possible gauge control systems cannot be stated in less than firm figures which show the increase in mill yield achieved by gauge control. Except for single-stand mills as recently developed in the United Kingdom, we are aware of no conclusive data in this matter.

References

1. Transfer Functions of Multi-Variable Machines. Part I. Tandem Rolling Mills. J. H. Courcoulas, J. M. Ham. *Research Report No. 6*, Control Systems Laboratory, Department of Electrical Engineering, University of Toronto, Toronto, Ont., Canada.

2. A Graphical Solution to the Cold Rolling Problem When Tensions Are Applied to the Strip. H. Ford, G. Lianis. *Journal*, Institute of Metals, London, England, vol. 84, 1955–56, p. 299.

Feedback Techniques Applied to the Single-Phase Transformer

E. MISHKIN
ASSOCIATE MEMBER AIEE

THE ANALYSIS of electric machines makes use of the vector diagram and the equivalent circuit diagram. Both of these methods give the various quantities sought (current, voltage, torque, etc.) under conditions of steady-state operation. When electric machines are used as elements in a general control system, one should rather apply the same techniques used for system analysis to the performance calculation of the particular element. In this case one is more interested in the machine's dynamic behavior, lag or lead it introduces, band width, frequency response, and an easy method for computing the modification or compensation it causes.

The block diagram method used in feedback control system could be applied advantageously to the analysis of electric machines as it would meet the aforementioned requirements and would divulge any closed loops inherent in the machine. It presents a better model of the machine's physical operation than the equivalent circuit diagram and is actually a generalization of the vector diagram, the latter presenting a single point on the well-known KG locus.

A block diagram representation of the single-phase transformer and the solution under various load and input conditions is given in the following. This can be adapted to both induction and synchronous machines. The saturation and hysteresis effects have been neglected in order to obtain a linear system.

The actual transformer is shown in Fig. 1(A) while its presentation as a feedback system is given in Fig. 1(B). Let V_1 be the applied voltage and E_1 and E_{L1} the feed-back electromotive force (emf) due to rate of change of main flux Φ and the primary leakage flux Φ_{L1}. (V_1 and all other following quantities are the Laplace transforms of the time variable quantities.) The "error voltage" $V_1+E_{L1}+E_1$ causes the primary current I_1 to flow through the resistance of the primary winding. The sum of the primary ampere-turns N_1I_1 and the secondary ampere-turns N_2I_2 are the magnetizing ampere-turns N_1I_μ. This "error current" N_1I_μ produces in turns the main magnetic flux Φ ($\mathcal{Q}$ denotes the reluctance of magnetic circuit) and the primary emf E_1. Let N_1/N_2 be the turns ratio, V_2 the secondary voltage, and E_{L2} the back emf due to secondary leakage flux Φ_{L2}. The secondary error voltage $E_2-V_2+E_{L2}$ is the ohmic voltage drop I_2R_2, where R_2 is the resistance of the secondary winding. Z denotes the transformer's load impedance.

A few limiting cases of the transformer's operation will be considered.

In the case of a d-c supply $s=0$ and the whole feedback system breaks down.

Under no-load conditions Z is to be assumed infinite. A finite secondary voltage V_2 postulates zero secondary current I_2. Consequently no I_2 is being fed back. The simplified loop diagram shown in Fig. 2 yields directly:

$$\frac{V_2'}{V_1}=\frac{-\frac{sN_1^2}{\mathcal{R}}}{R_1+sL_1+\frac{sN_1^2}{\mathcal{R}}} \tag{1}$$

where $V_2'=N_1/N_2$, V_2 is the secondary voltage referred to the primary since

$$\frac{N_1I_\mu}{\mathcal{R}}=\Phi \tag{2}$$

one can demote

$$\frac{N_1^2}{\mathcal{R}}=\frac{N_1\Phi}{I_\mu}=L \tag{3}$$

hence

$$\frac{V_2'}{V_1}=\frac{-sL}{R_1+s(L+L_1)} \tag{4}$$

$$\frac{I_\mu}{V_1}=\frac{1}{R_1+s(L+L_1)} \tag{5}$$

Fig. 3 shows the voltage ratio $|V_2'/V_1$ in decibels and the phase angle of V_2'/V_1 versus the angular frequency.

When the frequency of the applied supply is very low, the whole applied voltage V_1 is wasted in the primary ohmic drop and only a very small emf E_1 is fed back. The primary current is consequently very high while the secondary voltage is vanishing.

Paper 56-791, recommended by the AIEE Feedback Control Systems Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Summer and Pacific General Meeting, San Francisco, Calif., June 25–29, 1956. Manuscript submitted September 14, 1954; made available for printing May 10, 1956.

E. Mishkin is with the Polytechnic Institute of Brooklyn, Brooklyn, N. Y.

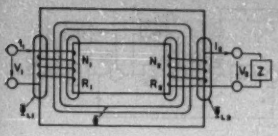


Fig. 1(A) (left). Elementary transformer

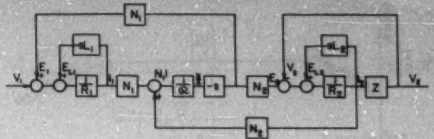


Fig. 1(B) (right). Feedback diagram of a transformer

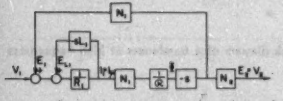


Fig. 2. No-load diagram

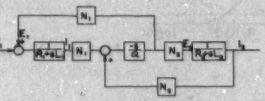


Fig. 4. Short-circuit diagram

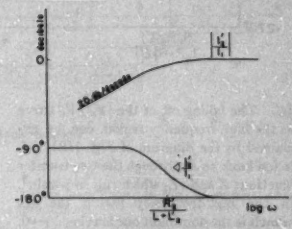


Fig. 5. Currents ratio (in decibels) and phase angle versus log ω under short-circuit conditions

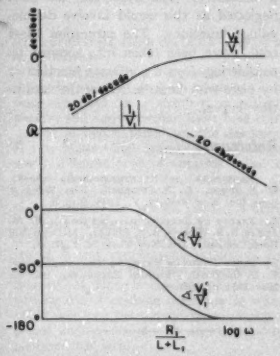


Fig. 3. No-load $|V_2'/V_1|$ and $|I_1/V_1|$ (in decibels) and their phase angles versus log ω

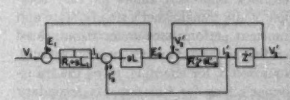


Fig. 6. Feedback diagram of a transformer

A short circuit occurs when $Z=0$. Fig. 1(B) can now be redrawn as shown in Fig. 4. From the diagram one obtains directly:

$$\frac{I_1}{V_1}=\frac{R_2'+s(L+L_2')}{R_1R_2'+s[R_1L_2'+R_2'L_1+(R_1+R_2')L]+s^2[L_1L_2'+L(L_1+L_2')]} \tag{6}$$

where R_2' and L_2' are the secondary resistance and leakage reactance respectively referred to the primary, i.e.,

$$R_2'=\left(\frac{N_1}{N_2}\right)^2 R_2;\quad L_2'=\left(\frac{N_1}{N_2}\right) L_2 \tag{7}$$

$$\frac{I_2'}{V_1}=\frac{-sL}{R_1R_2'+s[R_1L_2'+R_2'L_1+(R_1+R_2')L]+s^2[L_1L_2'+L(L_1+L_2')]};$$

$$I_2'=\frac{N_2}{N_1}I_2 \tag{8}$$

$$\frac{I_2'}{I_1}=\frac{-sL}{R_2'+s(L+L_2')} \tag{9}$$

Usually $L_1\cong L_2'<<L$ and equations 6 and 8 take the form

$$\frac{I_1}{V_1}=\frac{\dfrac{1}{R_1}+s\dfrac{L}{R_1R_2'}}{1+sL\left(\dfrac{1}{R_1}+\dfrac{1}{R_2'}\right)+s^2L\dfrac{L_1+L_2'}{R_1R_2'}} \tag{10}$$

$$\frac{I_2'}{V_1}=\frac{-s\dfrac{L}{R_1+R_2'}}{1+sL\left(\dfrac{1}{R_1}+\dfrac{1}{R_2'}\right)+s^2L\dfrac{L_1+L_2'}{R_1R_2'}} \tag{11}$$

The denominator in the last equations is the well-known quadratic

$$\frac{s^2}{\omega_n^2}+2\zeta\frac{s}{\omega_n}+1$$

where ω_n is the resonant angular frequency and ζ the damping factor. Numerical values taken from actually built power transformers show that ω_n is well below the usual supply frequency and that $\zeta>>1$; i.e., the system is highly overdamped. Fig. 5 shows the ratios $|I_2'/I_1|$ in decibels and the corresponding phase angle versus the frequency.

Normal Operation

The block diagram, Fig. 1(B), can be retraced as shown in Fig. 6, from which one obtains

$$\frac{V_2'}{V_1}=\frac{-sLZ'}{R_1(R_2'+Z')+s[R_1L_2'+(R_2'+Z')L_1+(R_1+R_2'+Z')L]+s^2[L_1L_2'+L(L_1+L_2')]} \tag{12}$$

$$\frac{I_1}{V_1}=\frac{R_2'+Z'+s(L+L_2')}{R_1(R_2'+Z')+s[R_1L_2'+(R_2'+Z')L_1+(R_1+R_2'+Z')L]+s^2[L_1L_2'+L(L_1+L_2')]} \tag{13}$$

(For methods of solving feedback block diagram see references 1, 2, and 3.)

$$\frac{I_2'}{I_1}=\frac{-sL}{R_2'+Z'+s(L+L_2')} \tag{14}$$

In case of an ideal transformer $R_1=R_2'=L_1=L_2'=0$ and $L=\infty$ and the foregoing equations yield the well-known formulas

$$\frac{V_2'}{V_1}=\frac{I_1}{I_2'}=-1 \tag{15}$$

Usually $L_1\cong L_2'<<L$ and equation 12 takes the form

$$\frac{V_2'}{V_1}=\frac{-sLZ'}{R_1(R_2'+Z')+sL(R_1+R_2'+Z')+s^2L(L_1+L_2')} \tag{16}$$

For $Z'=R_L'+sL_L'$ the last equation reads

$$\frac{V_2'}{V_1}=\frac{-sLR_2'-s^2LL_L'}{R_1(R_2'+R_L')+s[L(R_1+R_2'+R_L')+R_1L_L']+s^2L(L_1+L_L'+L_L')} \tag{17}$$

and in a similar way one can obtain the transformer's transfer function in case of a mixed resistance-inductance-capacitance load. The corresponding frequency response is obtained after inserting $s=j\omega$. Fig. 7 shows the ratio $|V_2'/V_1|$ (in decibels) and the phase angle of V_2'/V_1 versus the angular frequency for a purely ohmic load. It is clear from it that in order to increase the flat part of the $|V_2'/V_1|$ curve in the high-frequency region one has to decrease the total leakage L_1+

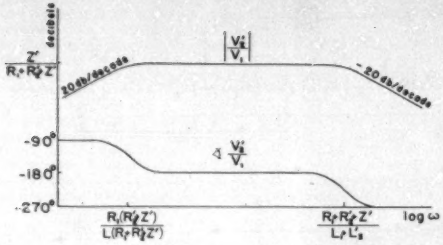


Fig. 7 (left). Voltages ratio (in decibels) and phase angle versus log ψ

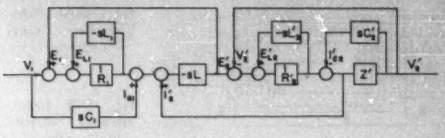


Fig. 8 (above). Feedback diagram of a transformer at high frequencies

L_2'. The falling off of the $|V_2'/V_1|$ curve in the high-frequency region can be explained by the diagram of Fig. 1(B). I_2 is fed back to E_2 through the two transfer functions Z and sL_2 which run in parallel. In the high-frequency region the sL_2 branch is the dominant one leaving $V_2 \to 0$. It is worth noting that the phase angle of V_2'/V_1 in this region tends to 270 degrees.

The foregoing analysis based on the feedback approach allows an easy solution of the transformer's steady-state and transient performance under various load conditions and various supply sources. Equation 5, for example, gives the transformer's transfer function, its secondary being open-circuited. The transformer's primary is assumed to be connected, at $t=0$, to a sine voltage source:

$$V_1(t) = V_{1m} \sin(\omega t + \varphi) \tag{18}$$

The Laplace transform of equation 18 is

$$V_1(s) = V_{1m} \frac{\omega \cos\varphi + s \sin\varphi}{s^2+\omega^2} \tag{19}$$

and the magnetizing current, in accordance with equation 2, will be

$$I_\mu(s) = \frac{\frac{1}{R_1}}{\tau s+1} V_{1m} \frac{\omega \cos\varphi + s \sin\varphi}{s^2+\omega^2}$$

$$\tau = \frac{L_1+L}{R_1} \tag{20}$$

The time dependence of the magnetizing current I_μ is given by the inverse Laplace transform of equation 20

$$I_\mu(t) = \frac{V_{1m}}{R_1(\omega^2\tau^2+1)} [\sin(\omega t+\varphi) - \tau\omega\cos(\omega t+\varphi) + (\tau\omega\cos\varphi - \sin\varphi)e^{-t/\tau}] \tag{21}$$

The first two terms in the last equation constitute the steady-state no-load current, while the last term expresses the transient contribution to this current. Its dependence on the phase φ of the supply voltage at the instant of switching in is clear.

The foregoing equations will hold for lower and audio frequencies. In case of higher frequencies one cannot neglect the distributed capacities between the individual winding turns. A fraction only of the displacement current of these capacities will flow through the windings, the rest of the current being in parallel with the windings. As far as the transformer's performance is concerned, only the current through the windings is to be considered. The rest of it represents an additional (capacitive) load on the applied primary voltage or can be lumped with the connected load. The corrected block diagram of the transformer is shown in Fig. 8, C_1 being the lumped capacity of the primary and C_2' that of the secondary, referred to the primary. The transformer's transfer function derived from this block diagram is (considering $L_1 \cong L_2' << L$):

$$\frac{V_2'}{V_1} = \frac{-sLZ'(1+sR_1C_1+s^2L_1C_1)}{R_1(R_2'+Z')+s[L(R_1+R_2'+Z')+Z'R_1R_2'C_2']+s^2L(L_1+L_2'+Z'R_2'C_2')+s^3LL_2'Z'C_2'} \tag{22}$$

where Z' is the actual load shunted by a fraction of the secondary windings lumped capacity. It is to be noted that contrary to the curve $|V_2'/V_1|$ shown in Fig. 7, the ratio $|V_2'/V_1|$ given by equation 22 tends to a certain constant value for very high frequencies.

Conclusion

The feedback approach presented in the previous paragraphs is only another method of analyzing the transformer's performance. By the use of a few principles of feedback engineering and the Laplace transform model is obtained which is suitable for computation of the transformer's static as well as dynamic behavior. The method is suitable for purpose of calculations and furnishes a direct physical interpretation. It combines thereby merits of both the vector diagram and the equivalent circuit diagram methods. It is clear that this approach could be profitably used in solving many other problems.

In all the aforementioned block diagrams R_1 could comprise the eddy-current effect if necessary. The hysteresis has been neglected as this would involve double-valued functions. The saturation effect could have been taken into account by considering L as a nonlinear function or by piece-wise linearization of the saturation curve.

References

1. Principles of Servomechanisms (book), G. S. Brown, D. P. Campbell. John Wiley & Sons, Inc., New York, N. Y., 1948, chap. 5.

2. Theory of Servomechanisms (book), H. M. James, N. B. Nichols, R. S. Phillips. McGraw-Hill Book Company, Inc., New York, N. Y., 1947.

3. Block Diagram Network Transformation, T. D. Graybeal. *Electrical Engineering*, vol. 70, Nov. 1951, pp. 985–90.

Discussion

T. M. Stout (Ramo-Wooldridge Corporation, Los Angeles, Calif.): Professor Mishkin's paper provides an interesting illustration of the utility of block diagrams, originally considered as a tool for analysis of feedback control systems, in the study of electric circuits and machines. In addition to requesting clarification of several points in the paper, I wish to call attention to several other papers in this same area and to suggest some extensions of the ideas discussed here.

In Fig. 8, a current I_{c1} is derived directly from voltage V_1 by multiplying by sC_1. This implies that the corresponding capacitor is connected directly across the source of voltage V_1. If so, it is not dependent on the transformer quantities and cannot affect the output voltage, as indicated by the block diagram. The effect of a capacitor shunting the secondary, in parallel with the load, is represented by C_2'; it is not clear why this should be described as a secondary capacity referred to the primary.

Although the author states in the text that R_1 could represent the eddy current losses of the transformer, it is more accurate and customary to represent these losses by a shunt resistor.[1] Treatment of these losses, as well as stray capacitance, would be facilitated by prior construction of an equiva-

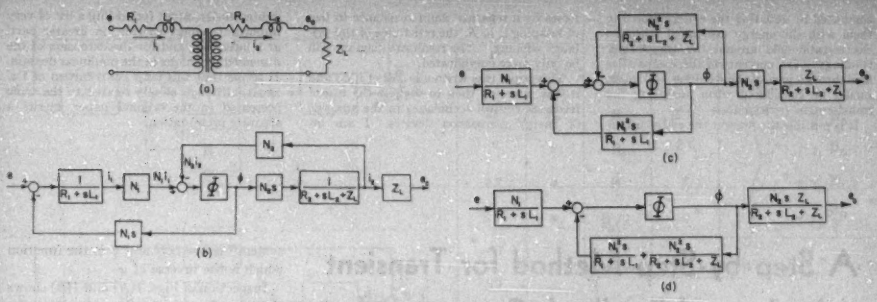


Fig. 9. Transformer with a nonlinear core material and block diagrams

lent circuit from which the block diagram is to be derived.

The analysis of electric circuits and machines by means of block diagrams has now been discussed in a number of papers. Passive and active circuits have been considered in two papers[1,2] by the discusser, and by Mason[4] and Truxal,[5] both of whom use a different type of notation called a "signal flow graph." Similar notation is used by Tustin[6] in a thorough treatment of d-c machines. In a work published after Professor Mishkin's paper was submitted, block diagrams have been applied to d-c machines and synchronous generators by Riaz,[7,8] and to magnetic amplifiers by Storm.[9] In fact, block diagrams are appearing with increasing frequency in what might once have been considered unexpected places, and this list of references is probably far from complete.

Block diagram representation of systems containing nonlinear relations is a problem in which I have been interested for some time. The block diagram for a transformer with a nonlinear core material, taken from a recent paper,[10] is shown in Fig. 9. The nonlinear relationship between net ampere-turns and flux is represented by Φ. The similarity of these diagrams to those in the paper is evident.

It is easy to show[10] that all circuits containing a single nonlinear element characterized by a function $y=f(x)$, with a single input and output, can be represented by a diagram having the general form given in Fig. 9(d) plus a linear feed-forward path direct from the input to the output. Similar but more complicated general diagrams can be drawn[10] for systems containing two nonlinear elements, $y=f(x)$ and $u=g(v)$, or a single nonlinear relation between three variables, $z=f(x,y)$.

A method for transient analysis of systems containing a nonlinear element was recently presented.[11] The method employs operational calculus, numerical integration, and graphical solution of a linear algebraic equation with the nonlinear characteristic, used in its original form without piece-wise linearization. Frequency-response analyses can be performed by replacing the characteristic curve by a describing function or equivalent impedance.[12,13]

Qualitative insights into system behavior can be gained without extensive analysis. For example, the response of the transformer to large sinusoidal driving voltages might be desired as a function of the load impedance magnitude. Feedback theory[14] states that distortion produced by nonlinearity is reduced most effectively when the loop gain is high. Since Z_L appears in the denominator of the feedback quantity in Fig. 9(d), the loop gain will vary inversely with Z_L. The distortion should therefore be greatest when Z_L is large, i.e., when the secondary is open-circuited. Moreover, the diagram also indicates that the magnitude of the output voltage decreases with Z_L. Experimental data for a small transformer (specifications for which are no longer available) with a resistive load, shown in Fig. 10, support these predictions.

Fig. 10. Response of transformer in Fig. 9 to a sinusoidal voltage of fixed amplitude and frequency for various load resistances

Open circuit
5,000 ohms
2,200 ohms
1,250 ohms
600 ohms

References

1. Magnetic Circuits and Transformers (book), Electrical Engineering Staff. Massachusetts Institute of Technology. John Wiley & Sons, Inc., New York, N. Y., 1947.

2. A Block-Diagram Approach to Network Analysis, T. M. Stout. *AIEE Transactions*, vol. 71, pt. II, 1952, pp. 255–60.

3. Block Diagram Solutions for Vacuum-Tube Circuits, T. M. Stout. *Ibid.*, vol. 72, pt. I, Nov. 1953, pp. 361–67.

4. Feedback Theory—Some Properties of Signal Flow Graphs, S. J. Mason. *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 41, Sept. 1953, pp. 1144–56.

5. Automatic Feedback Control System Synthesis (book), J. G. Truxal. McGraw-Hill Book Company, Inc., New York, N. Y., 1955, pp. 88–160.

6. Direct Current Machines for Control Systems (book), A. Tustin. The MacMillan Company, New York, N. Y., 1952.

7. Dynamics of D-C Machine Systems, M. Riaz. *AIEE Transactions*, vol. 74, pt. II, 1955 (Jan. 1956 section), pp. 365–70.

8. Analogous Computer Representations of Synchronous Generators in Voltage Regulation Studies, M. Riaz. *Ibid.*, vol. 75, pt. III, Dec. 1956, pp. 1178–84.

9. Magnetic Amplifiers (book), H. F. Storm. John Wiley & Sons, Inc., New York, N. Y., 1955.

10. Block Diagram Transformations for Systems with One Nonlinear Element, T. M. Stout. *AIEE Transactions*, vol. 75, pt. II, July 1956, pp. 130–41.

11. A Step-by-Step Method for Transient Analysis of Feedback Systems with One Nonlinear Element, T. M. Stout. *Ibid.*, vol. 75, pt. II, 1956 (Jan. 1957 section), pp. 378–90.

12. A Frequency-Response Method for Analyzing and Synthesizing Contactor Servomechanisms, Ralph J. Kochenburger. *Ibid.*, vol. 69, pt. I, 1950, pp. 270–84.

13. A Method of Approximate Steady-State Analysis of Nonlinear Networks, G. R. Slemon. *Proceedings*, Institution of Electrical Engineers, London, England, vol. 100, pt. I, Sept. 1953, pp. 275–87.

14. Network Analysis and Feedback Amplifier Design (book), H. W. Bode. D. Van Nostrand Company, Inc., New York, N. Y., 1945.

E. Mishkin: One can substitute lumped capacities on the primary and secondary sides of the transformer representing the actually distributed capacities between the individual turns of the windings, the turns and the core, etc. They can be

computed by equating the energy stored in them with the energy associated with the electrostatic field around the transformer windings. The currents in these capacities flow through part of the windings and hence should be considered when analyzing the transformers performance.

It is possible the present the eddy current losses by a separate shunt resistance in lieu of lumping it in R, the resistance of the primary winding. The resultant diagram will be only more complicated.

This paper was written in fall of 1953 and it was intended then to suggest the use of feedback control techniques in the analysis of energy conversion devices. I am indebted to Dr. Stout for adding a list of very important papers written, in greater part, at a later date and for the extension of the discussed technique in the nonlinear domain. It seems that this later contribution of Dr. Stout, although closely related to the topic presented in the original paper, merits a separate presentation.

A Step-by-Step Method for Transient Analysis of Feedback Systems with One Nonlinear Element

T. M. STOUT
ASSOCIATE MEMBER AIEE

Synopsis: An analytical-numerical-graphical method is presented for the transient analysis of nonlinear feedback systems. The method is applicable to systems of any number of linear elements and one nonlinear element whose characteristic may be represented by $y=\varphi(x)$, where x and y are its instantaneous input and output. The techniques employed are operational mathematics, numerical integration, and a graphical solution of two algebraic relations. Several examples, including both electric networks and a feedback control system, illustrate the procedure. Experimental results and theoretical considerations relating to the accuracy of the method are given.

NONLINEARITY is sometimes useful, sometimes a nuisance, and sometimes merely interesting. In many cases, nonlinearity is merely a slight complication and linear differential equations with constant coefficients, which can be solved by routine methods taught in elementary courses in differential equations or electrical transients, provide an adequate approximation. In other cases, phenomena occur which defy explanation unless relevant nonlinearities are considered. In electrical engineering alone, examples which might be listed include subharmonic resonance and jump phenomena in circuits containing iron-cored reactors; generation of harmonics, subharmonics, and intermodulation products in communication systems; rectification and detection; amplification by means of nonlinear magnetic or dielectric materials; amplitude limiting in vacuum-tube oscillators; relaxation oscillations; and hunting in automatic control systems or systems of coupled synchronous machines.

The mathematical theory available for the treatment of such problems is, understandably, in an unsatisfactory state; there is no standard method by which all nonlinear differential equations can be solved. A limited number of these equations can be solved exactly,[1-3] and an almost bewildering variety of approximate analytical, numerical, and graphical methods are available for the rest.[4-27] These methods are not generally interchangeable, and each has its advantages and disadvantages for particular purposes.

The present method is applicable to systems having one input, one output, and one nonlinear element which can be characterized by a relation between its instantaneous input and output, say $y=\varphi(x)$. Such systems can always be represented by the block diagrams of Figs. 1(A) and 1(B).[30] In these diagrams, G_1, G_2, G_3, G, and H are transfer functions which represent linear dynamic relations (also expressible by linear differential equations) and φ and ψ represent functions relating x and y. The two diagrams are interchangeable descriptions of the same system[30] if $G=1/H$ and ψ is the function which is the inverse of φ.

Inspection of Figs. 1(A) and 1(B) shows that u, a, and b are the outputs of the linear blocks G_1, G_2, and G_3, and can easily be computed, by any of several different methods, when the inputs are known. The crucial problem is evidently the analysis of the nonlinear feedback system composed of φ and H. In the proposed method, the value of v at any time is expressed as a weighted sum of present and past values of y by means of a numerical approximation of the superposition integral relating v and y; the weights depend on the impulse response of the block H and on the interval chosen. A graphical solution solution of two algebraic equations is required at each step to find the present values of x, y, and v.

The proposed method is somewhat similar to methods previously described by Tustin[28,29] Madwed,[27] Wagner,[26] and Truxal.[25] Unlike these other methods, this method does not require resolution of the variables into step, pulse, or other component signals, or the substitution of difference for differential operators; no attempt is made to use regression formulas or Z-transform techniques. The weighting function required in the analysis is determined by inverse Laplace transformation of the appropriate transfer function, and the superposition integral is evaluated by standard numerical methods. If this approach seems lacking in elegance, it has the advantage of using only mathematical techniques which are generally familiar to engineers. Most important, use of standard numerical integration procedures suggests the use of associated formulas for the estimation of errors, with the result that the accuracy of the results can be estimated and a suitable step size may be selected after one set of trial calculations.

After an explanation of the method, three nonlinear electric circuits and a saturating feedback control system are analyzed and checked experimentally. A theoretical discussion of the accuracy of the method concludes the paper.

Paper 56-779, recommended by the AIEE Feedback Control Systems Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Summer and Pacific General Meeting, San Francisco, Calif., June 25-29, 1956. Manuscript submitted October 13, 1955; made available for printing April 27, 1956.

T. M. STOUT is with the Ramo-Wooldridge Corporation, Los Angeles, Calif.

This paper is based on work done at the University of Washington, Seattle, Wash., under contract with the Office of Naval Research and included in a thesis submitted to the University of Michigan, Ann Arbor, Mich., in partial fulfillment of requirements for the doctor of philosophy degree.

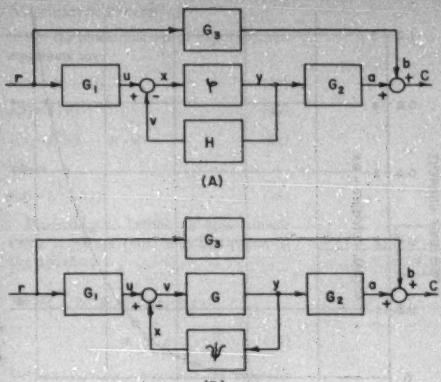


Fig. 1 (left). Standard block diagrams for systems with one nonlinear element

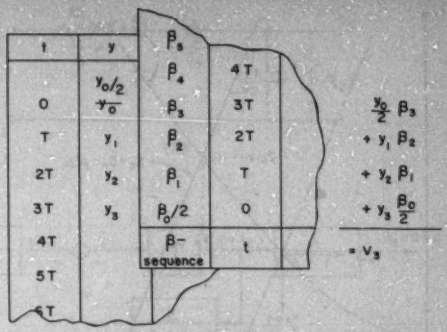


Fig. 2 (above). Approximate evaluation of the superposition integral

Description of Analytical Method

Since the input block G_1 is linear, its output $u(t)$ can be calculated for any given input $r(t)$ by any of several methods:

1. The operational expression

$$U(s) = G_1(s)R(s) \tag{1}$$

can be converted to an ordinary differential equation and solved by classical methods.

2. Laplace transform methods can be applied directly, giving

$$u(t) = L^{-1}[G_1(s)R(s)] \tag{2}$$

where

$$R(s) = L[r(t)] \tag{3}$$

3. Time-domain methods, based on the superposition integral[21]

$$u(t) = \int_0^t G_1(t-\tau)r(\tau)d\tau \tag{4}$$

where

$$G_1(t) = L^{-1}[G_1(s)] \tag{5}$$

can be used, either exactly or approximately.

4. Frequency-response methods, based on the Fourier series or integral, can be used.

The same methods can be used to find $b(t)$ from $r(t)$ and $a(t)$ from $y(t)$. However, it should be noted that $y(t)$ is the output of a nonlinear feedback system and is not available in analytical form because of the approximate calculation procedure employed. For this reason, it will be necessary to use approximate methods even in this linear part of the system. As will be observed in the examples, this difficulty is sometimes mitigated by the fact that $a(t)$ is identical with, or simply related to, $v(t)$.

The real problem, with which this paper is principally concerned, is the calculation of the response of a nonlinear feedback system for fairly general inputs. Two similar procedures for this purpose will be described, one applicable to Fig. 1(A) which will be called the general procedure and another applicable to Fig. 1(B) called an alternative procedure.

General Procedure

The nonlinear feedback system of Fig. 1(A) may be described by the equations

$$x(t) = u(t) - v(t) \tag{6}$$

$$y(t) = \varphi\{x(t)\} \tag{7}$$

$$v(t) = \int_0^t h(t-\tau)y(\tau)d\tau \tag{8}$$

where

$$h(t) = L^{-1}[H(s)] \tag{9}$$

This description is entirely in the time domain and takes advantage of the fact that the output of a linear system can be expressed as a weighted sum of its present and past inputs. The weighting function, $h(t)$, is the impulse response of the feedback block (or the derivative of its step function response)[21] and may be calculated in advance, as can $u(t)$.

Equations 6–9 can be combined into a single equation

$$x(t) = u(t) - \int_0^t h(t-\tau)\varphi\{x(\tau)\}d\tau \tag{10}$$

in which $x(t)$ is the only unknown; when $x(t)$ is found, $y(t)$ is available from equation 7. Equation 10 is a nonlinear integral equation, other versions of which have been applied to electric circuit problems by Carson,[17] Keller,[18] and Pipes.[19] Solutions of this equation can be found by an iterative procedure, using $u(t)$ as a first approximation to $x(t)$ which in turn is used to find a second approximation, and so on. This procedure requires an analytic expression for $\varphi(x)$ and the integrands increase rapidly in complexity, even under the best conditions.

An approximate solution can be found by step-by-step methods which permit use of $\varphi(x)$ in graphical form and lead directly to numerical results. The approach employed here, evaluation of the superposition integral by standard methods of numerical integration, has been used by Collatz[23] and Wagner[20] for linear integral equations similar to equation 10.

Using numerical methods, the various system variables are computed at intervals $\Delta t = T$; these values will be denoted by $v_n = v(nT)$, $h_n = h(nT)$, and so forth. If the trapezoidal rule[19] is used, the approximate value of $v(nT)$ from equation 8 becomes

$$v_n \cong T\left[\frac{1}{2}h_n y_0 + h_{n-1}y_1 + h_{n-2}y_2 + \ldots + h_2 y_{n-2} + h_1 y_{n-1} + \frac{1}{2}h_0 y_n\right] \tag{11}$$

More accurate values of v_n could be obtained by using more elaborate integration formulas, but none of these formulas are as convenient as the trapezoidal rule. If one writes

$$\beta_n = Th_n \tag{12}$$

then equation 11 can be regarded as the product of the two sequences

$$\frac{y_0}{2}, y_1, y_2, y_3, \ldots, y_n$$

$$\frac{\beta_0}{2}, \beta_1, \beta_2, \beta_3, \ldots, \beta_n \tag{13}$$

according to the scheme shown in Fig. 2. The first value of $y(t)$, y_0, is found from the

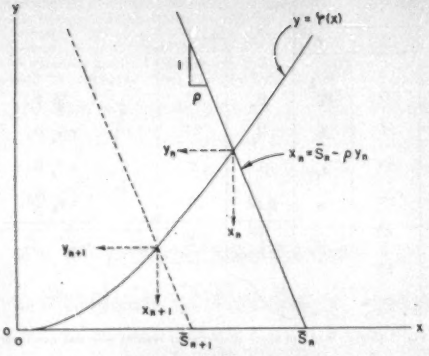


Fig. 3 (above). Graphical solution for x_n and y_n.

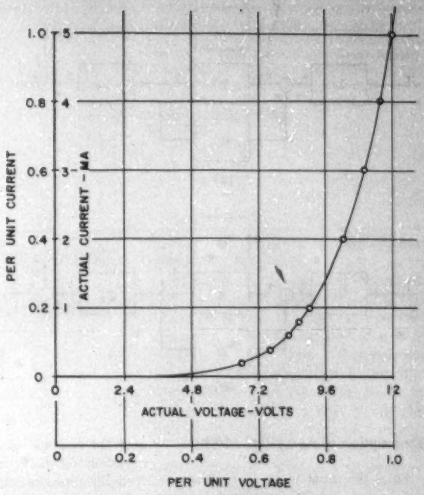


Fig. 4 (right). Diode characteristic

initial conditions of the problem; the rest of the values are determined step by step during the calculations. The calculations indicated in Fig. 2 are conveniently carried out on a desk calculator, recording only the final sum v_n at each step.

Unless $\beta_0 = 0$, it will be necessary to use a graphical solution of two algebraic equations at each step in order to determine v_n, x_n, and y_n. This complication is introduced by the necessity of satisfying the equations

$$y_n = \varphi(x_n) \tag{14}$$

$$x_n = u_n - v_n \tag{15}$$

simultaneously with equation 11. Letting S_n be the sum of all terms in equation 11 but the last, one may write

$$v_n = \rho y_n + S_n \tag{16}$$

where

$$\rho = \beta_0/2 \tag{17}$$

and

$$S_n = \frac{y_0}{2}\beta_n + y_1\beta_{n-1} + \ldots + y_{n-1}\beta_1 \tag{18}$$

Equation 15 may therefore be written

$$x_n = u_n - \rho y_n - S_n \tag{19}$$

$$= \bar{S}_n - \rho y_n \tag{20}$$

where

$$\bar{S}_n = u_n - S_n \tag{21}$$

Since u_n may be calculated in advance and S_n involves only past values of y, up to y_{n-1}, $\bar{S}_n$ is available at each step. The only unknown quantities are the present values of x and y, namely x_n and y_n, and these can be found by simultaneous solution of equations 14 and 20 using the graphical procedure of Fig. 3.

The graphical solution required is exactly the same as that used for the large-signal analysis of vacuum tubes with resistive loads, with x, y, and $\bar{S}$ analogous to e_b, i_b, and E_{bb}, respectively, and ρ analogous to R_L. The intercept $\bar{S}$ varies during the course of the calculations but the slope $1/\rho$ is constant. The construction is conveniently carried out by drawing a line of the proper slope on a piece of semitransparent tracing paper and superimposing this sheet on a plot of $\varphi(x)$, shifting it back and forth to the $\bar{S}$ required at each step.

Remarks

If $\beta_0 = \rho = 0$, the graphical solution is unnecessary. In this case, v_n depends only on past values of y, up to y_{n-1}, and can be found directly at each stage of the calculations. With v_n known, x_n is immediately available and y_n can be read from the curve.

Because $H(s)$ is normally the ratio of two polynomials in s, the relative orders of its numerator and denominator are of interest. In passive network problems, the orders may differ by as much as two; in active network and feedback control system problems, the orders may differ by more than two.

If the order of the denominator is greater than the order of the numerator, the procedure is straightforward; the inverse transform, $h(t)$, can be found by partial fraction expansion or by reference to a table of transforms.[31] Where the numerator and denominator are of the same order, it is advisable first to remove a constant term by division, as in the case

$$H(s) = \frac{2s^2+3s+1}{s^2+s+2}$$

$$= 2 + \frac{s-3}{s^2+s+2}$$

In this case a term $2y_n$ would be added to equation 20 to account for the constant term in $H(s)$ and the remaining fraction would be treated in regular fashion. In general, the added term would be Ky_n where K is the constant removed by division. The constant can also be regarded as the transfer function of a separate feedback block which may be combined with φ to produce a new nonlinear forward block.[30]

If the order of the numerator exceeds the order of the denominator, the inverse transform would contain impulsive components which would cause trouble in the analysis. This difficulty can be avoided by reversing the positions of the linear and nonlinear blocks, putting the reciprocal of $H(s)$ in the forward block and the inverse function for the nonlinearity into the feedback path.[30] This transformation gives Fig. 1(B) and requires a slightly different procedure.

Alternative Procedure

For the block diagram of Fig. 1(B), the equations are

$$v(t) = u(t) - x(t) \tag{22}$$

$$x(t) = \psi[y(t)] \tag{23}$$

$$y(t) = \int_0^t v(t-\tau) g(\tau) d\tau \tag{24}$$

where

$$g(t) = L^{-1}[G(s)] \tag{25}$$

Proceeding as before, an approximate value of $y(t)$ for $t = nT$ may be written as the equation

$$y_n \cong T\left[\frac{g_n}{2} v_0 + g_{n-1} v_1 + g_{n-2} v_2 + \ldots + g_2 v_{n-2} + g_1 v_{n-1} + \frac{g_0}{2} v_n\right] \tag{26}$$

Letting $\alpha_n = T g_n$, equation 26 may be regarded as the product of two sequences

$$\frac{v_0}{2}, v_1, v_2, v_3, \ldots, v_n \tag{27}$$

$$\frac{\alpha_0}{2}, \alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_n \tag{28}$$

calculated using the same scheme shown in Fig. 2.

In preparation for the graphical solution, equation 26 may be written

$$y_n = \mu v_n + Q_n \tag{29}$$

where

$$\mu = \alpha_0/2 \tag{30}$$

and

$$Q_n = \frac{v_0}{2}\alpha_n + v_1\alpha_{n-1} + \ldots + v_{n-1}\alpha_1 \tag{31}$$

Combining equations 22 and 29 results in the following

$$y_n = \mu(u_n - x_n) + Q_n \tag{32}$$

$$= \bar{Q}_n - \mu x_n \tag{33}$$

where

$$\bar{Q}_n = Q_n + \mu u_n \tag{34}$$

Equations 23 and 33 constitute a pair of simultaneous equations in x and y which may be solved by graphical process similar to that illustrated in Fig. 3.

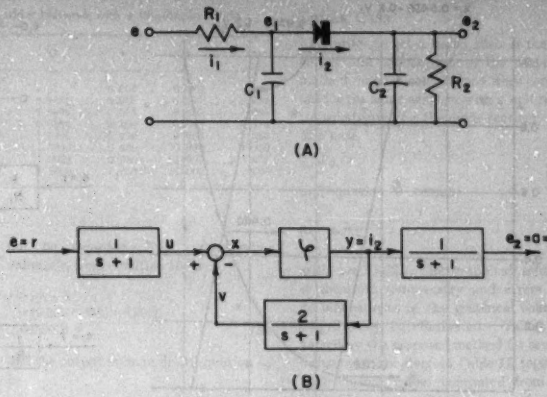


Fig. 5. 2-section R-C ladder network and block diagram

Frequency Response

No restriction has been placed on the nature of $r(t)$ and the method can, in principle, be used to find the steady-state response of a nonlinear system to sinusoidal inputs. Since the calculations start with the application of the sinusoidal input and continue until the response is periodic, this procedure entails a considerable amount of unproductive work if only the steady-state response is wanted. To avoid this extra labor, an extension of the describing function method[13] may be used.

The method, due to Cosgriff,[33] can be summarized as follows. A sinusoidal variation of x is assumed, having an amplitude X and a frequency ω. The magnitude and phase of the fundamental component of y, say Y, is found with the help of a describing function $F(X)$ which replaces $\varphi(x)$ as a description of the nonlinear element. The system variables are then

$$U(X, \omega) = X + H(\omega)F(X)X$$

$$R(X, \omega) = \frac{U(X, \omega)}{G_1(\omega)}$$

$$C(X, \omega) = G_2(\omega)F(X)X + G_2(\omega)R(X, \omega)$$

where all quantities have, in general, magnitude and phase. By assuming other amplitudes and frequencies, a complete description of the output-input ratio C/R as a function of frequency and input (or other) magnitude can be developed. A similar approach, using equivalent gains based on peak magnitudes instead of describing functions, is described by Prince.[35]

The accuracy of the results obtained by this method is increased if the harmonics actually present in y are effectively filtered by the linear parts of the systems. Methods for considering the effects of harmonics are found in papers by Johnson[34] and Slemon,[35] who discusses the analysis of nonlinear circuits using an equivalent impedance which serves the same purpose as the describing function.

Table I. Calculations for 2-Section R-C Ladder Network with a Nonlinear Diode

t	n	u	S	$\bar{S}$	y	v	e_2
0	0	0.0000			0.000	0.0000	0.0000
0.2	1	0.5438	0.0000	0.5438	0.038	0.0076	0.0038
0.4	2	0.9800	0.0124	0.9766	0.480	0.1084	0.0542
0.6	3	1.3536	0.1673	1.1863	0.964	0.3601	0.1800
0.8	4	1.6520	0.4577	1.1993	1.000	0.6527	0.3264
1.0	5	1.8964	0.6982	1.1982	1.000	0.8982	0.4491
1.2	6	2.0964	0.8992	1.1972	1.000	1.0992	0.5496
1.4	7	2.2602	1.0638	1.1964	1.000	1.2638	0.6319

Table II. Results of Calculations for 2-Section R-C Ladder Network with Linear Resistor

	Output Voltage			
	Approximate			
t	T = 0.10	T = 0.20	T = 0.40	Exact
0	0.0000	0.0000	0.0000	0.0000
0.1	0.0130			0.0132
0.2	0.0461	0.0453		0.0463
0.3	0.0918			0.0921
0.4	0.1449	0.1442	0.1413	0.1451
0.5	0.2016			0.2018
0.6	0.2594	0.2590		0.2594
0.7	0.3163			0.3164
0.8	0.3713	0.3714	0.3713	0.3714
0.9	0.4238			0.4238
1.0	0.4732	0.4735		0.4731
1.2		0.5625	0.5643	0.5619
1.4		0.6384		0.6378
1.6		0.7022	0.7049	0.7013
1.8		0.7553		0.7543
2.0		0.7992	0.8024	0.7983

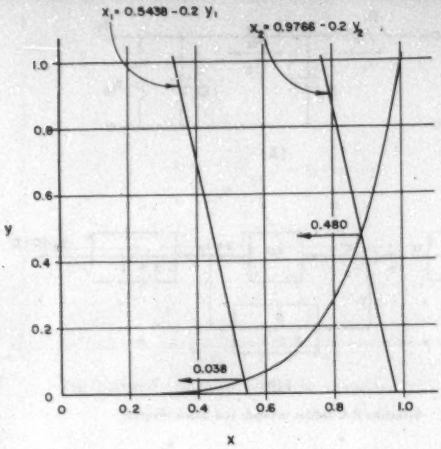


Fig. 6. Graphical solution for 2-section R-C ladder network

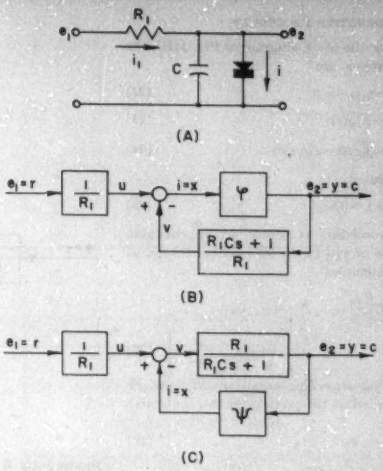


Fig. 8. 1-section R-C ladder network and block diagrams

Illustrative Examples

In the three electric networks which are used as examples, the nonlinear element consists of a series combination of four selenium rectifiers with the volt-ampere characteristic shown in Fig. 4. Voltages will be measured with respect to a base voltage of 12 volts and currents with respect to a base current of 5 milliamperes, giving a unit resistance of $12/0.005 = 2{,}400$ ohms. The unit of capacitance is 1 microfarad, making the time unit $(2{,}400)\ (10^{-6})$ seconds or 2.4 milliseconds. Element values are chosen to simplify the calculations.

Two-Section Resistance-Capacitance Ladder Network

The first example considered is the 2-section resistance-capacitance (R-C) ladder network shown in Fig. 5(A). The network parameters are $R_1 = R_2 = 2{,}400$ ohms, $C_1 = C_2 = 1$ microfarad; a 36-volt step will be applied at the input terminals of the network with the capacitors initially uncharged. On a per-unit basis, the values are $R_1 = R_2 = 1$, $C_1 = C_2 = 1$, and, $e = 3$. The block diagram for these particular values reduces to the form shown in Fig. 5(B).[20]

Preliminary Calculations

For the values selected, the preliminary calculations are as follows. The transform of the input voltage is

$$R(s) = E(s) = \frac{3}{s} \tag{35}$$

and the transfer function of the input block is

$$G_1(s) = \frac{1}{s+1} \tag{36}$$

so that

$$U(s) = \frac{3}{s(s+1)} \tag{37}$$

and

$$u(t) = 3(1 - e^{-t}) \tag{38}$$

The transfer function of the feedback block is

$$H(s) = \frac{2}{s+1} \tag{39}$$

so that

$$h(t) = 2e^{-t} \tag{40}$$

Taking $\Delta t = T = 0.48$ millisecond or $0.48/2.4 = 0.2$ unit, the β-sequence defined by equations 12 and 13 is

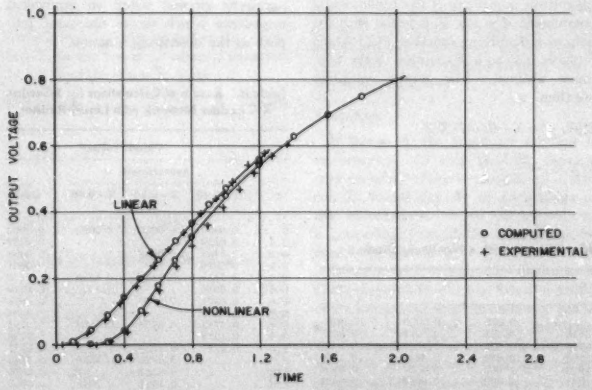


Fig. 7. Response of 2-section R-C ladder network

Table III. Calculations for 1-Section R-C Ladder Network with a Nonlinear Diode

t	n	u	μ0	Q	Q̄	x	y	v
0	0	2.0000	0.2000			0.000	0.000	1.000
0.2	1	2.0000	0.2000	0.1627	0.3637	0.010	0.362	1.990
0.4	2	2.0000	0.2000	0.4599	0.6599	0.095	0.650	1.906
0.6	3	2.0000	0.2000	0.6866	0.8866	0.380	0.850	1.620
0.8	4	2.0000	0.2000	0.8291	1.0291	0.750	0.950	1.250
1.0	5	2.0000	0.2000	0.8835	1.0835	0.950	0.990	1.050
1.2	6	2.0000	0.2000	0.8953	1.0953	0.980	1.000	1.020
1.4	7	2.0000	0.2000	0.9000	1.1000	1.000	1.000	1.000

$$\frac{0.4000}{2},\ 0.3274,\ 0.2682,\ 0.2196,\ \ldots \qquad (41)$$

and therefore $\rho=0.20$. The general solution for x_n is thus

$$x_n = S_n - 0.20 y_n \qquad (42)$$

For the particular values selected, it will be noted that

$$G_2(s) = \frac{H(s)}{2} \qquad (43)$$

so that

$$C(s) = E_2(s) = \frac{V(s)}{2} \qquad (44)$$

and

$$c(t) = e_2(t) = \frac{v(t)}{2} \qquad (45)$$

Transient Solution

Calculations are started by computing and tabulating values of u_n from equation 38, giving the first column of Table I. The initial conditions give

$$x_0 = y_0 = 0.0000 \qquad (46)$$

$$v_0 = 0.0000 \qquad (47)$$

At $t=0.20$ or $n=1$, from equations 18, 41, and 46,

$$S_1 = \left(\frac{y_0}{2}\right)(\beta_1)$$

$$= (0.0000)(0.3274)$$

$$= 0.0000 \qquad (48)$$

From equations 21, 38, and 48,

$$S_1 = u_1 - S_1$$

$$= 0.5438 \qquad (49)$$

Substituting equation 49 into equation 42 gives

$$x_1 = 0.5438 - 0.20 y_1 \qquad (50)$$

which is solved simultaneously with the diode characteristic, as shown in Fig. 6, giving

$$y_1 = 0.038 \qquad (51)$$

$$x_1 = 0.536 \qquad (52)$$

The value of y is entered in Table I; the values of x are not used and therefore need not be recorded. The corresponding value of v, from equation 16, is

$$v_1 = \rho y_1 + S_1$$

$$= (0.20)(0.038) + 0.0000$$

$$= 0.0076 \qquad (53)$$

and the output voltage, from equation 45, is

$$e_2(0.2) = \frac{v_1}{2}$$

$$= 0.0038 \qquad (54)$$

At $t=0.40$ or $n=2$,

$$S_2 = \beta_1 y_1 + \beta_2 \left(\frac{y_0}{2}\right)$$

$$= (0.3274)(0.038) + (0.2682)(0.0000)$$

$$= 0.0124 \qquad (55)$$

$$S_2 = u_2 - S_2$$

$$= 0.9890 - 0.0124$$

$$= 0.9766 \qquad (56)$$

$$x_2 = 0.9766 - 0.20 y_2 \qquad (57)$$

Simultaneous solution of equation 57 with the diode characteristic, shown in Fig. 6, gives

$$y_2 = 0.480 \qquad (58)$$

$$v_2 = (0.20)(0.480) + 0.0124$$

$$= 0.1084 \qquad (59)$$

$$e_2(0.4) = \frac{0.1084}{2}$$

$$= 0.0542 \qquad (60)$$

At $t=0.6$ or $n=3$,

$$S_3 = \beta_1 y_2 + \beta_2 y_1 + \beta_3 \left(\frac{y_0}{2}\right)$$

$$= (0.3274)(0.480) + (0.2682)(0.038) + (0.2196)(0.000)$$

$$= 0.1673 \qquad (61)$$

The remaining steps follow the same pattern as before and the results are given in Table I.

Similar calculations for $\Delta t = T = 0.1$ give a more complete picture of the early part of the transient but give values which differ only slightly from those in Table I, indicating that in this case the graphical solution limits the accuracy obtainable.

Linear Case

In order to get a better idea of the accuracy and convergence of the proposed method, similar calculations were carried out for the same network with a unit resistor substituted for the diode rectifier. In this case,

$$x_n = y_n \qquad (62)$$

and equation 20 becomes

$$x_n = y_n = \frac{S_n}{1+\rho} \qquad (63)$$

with $\rho = Th_0/2$. The graphical solution is therefore unnecessary and errors due to inaccuracies in the graphical solution can thereby be eliminated. Values obtained by the proposed method for several values of T are given in Table II, together with correct values computed from the equation

$$e_2(t) = 1 - \frac{3}{2}\,\varepsilon^{-t} + \frac{1}{2}\,\varepsilon^{-3t} \qquad (64)$$

obtained by ordinary operational methods.

Experimental Verification

Experimental results for both nonlinear and linear versions of this network, traced from photographs of a cathode-ray oscilloscope, are shown in Fig. 7. Agreement between computed and experimental results is satisfactory.

One-Section Resistance-Capacitance Ladder Network

The second example considered is the 1-section R-C ladder network shown in Fig. 8(A). The element values used are $R_1 = 2{,}400$ ohms and $C = 1$ microfarad; in per-unit values, $R_1 = 1$ and $C = 1$. A

Table IV. Results of Calculations for 1-Section R-C Ladder Network with a Linear Resistor

	Output Voltage			
	Approximate			
t	T = 0.10	T = 0.20	T = 0.40	Exact
0	0.0000	0.0000	0.0000	0.0000
0.1	0.1814			0.1813
0.2	0.3300	0.3306		0.3297
0.3	0.4516			0.4512
0.4	0.5511	0.5522	0.5568	0.5507
0.5	0.6326			0.6321
0.6	0.6993	0.7006		0.6988
0.7	0.7540			0.7534
0.8	0.7987	0.8001	0.8055	0.7981
0.9	0.8353			0.8347
1.0	0.8653	0.8666		0.8647
1.2		0.9112	0.9168	0.9098
1.4		0.9410		0.9392
1.6		0.9610	0.9665	0.9592
1.8		0.9743		0.9737
2.0		0.9832	0.9887	0.9817

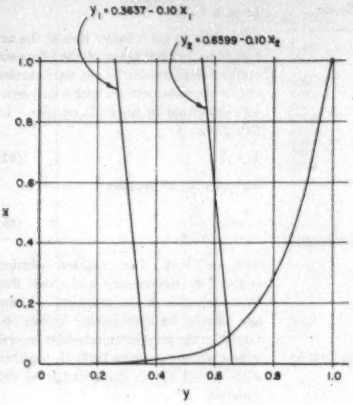


Fig. 9. Graphical solution for 1-section R-C ladder network

24-volt or 2-unit step will be applied at the input terminals of the network with the capacitor initially uncharged.

A possible block diagram for this network is shown in Fig. 8(B).[80] For the values chosen, the feedback block is

$$H(s) = s + 1$$

whose numerator is of higher order than the denominator. As has been indicated previously, it is necessary to interchange the forward and feedback blocks, giving the block diagram of Fig. 8(C). In this diagram, $x = \psi(y)$ represents the diode rectifier current as a function of the voltage; this function is given by the same curve, Fig. 4, used in the previous example.

Preliminary Calculations

In this particular problem, the input block has a transfer function

$$G_1(s) = 1 \tag{65}$$

so that $u(t)$ is numerically (but not dimensionally) equal to the applied voltage e_1 and therefore

$$u_n = i_n = 2.0000 \tag{66}$$

The transfer function of the forward block is

$$G(s) = \frac{1}{s+1} \tag{67}$$

so that

$$g(t) = e^{-t} \tag{68}$$

For $\Delta t = T = 0.20$ unit, the α-sequence defined by equations 27 and 28 is

$$\frac{0.2000}{2},\ 0.1637,\ 0.1098,\ 0.0899,\ \ldots \tag{69}$$

and $\mu = 0.10$ from equation 30. The special forms of equations 33 and 34 required for this problem are thus

$$y_n = \bar{Q}_n - 0.10 x_n \tag{70}$$

$$\bar{Q}_n = Q_n + 0.10 u_n \tag{71}$$

However, since $u_n = 2.0000$, equation 71 degenerates into

$$\bar{Q} = Q_n + 0.2000 \tag{72}$$

Transient Solution

The initial conditions give

$$y_0 = x_0 = 0.0000 \tag{73}$$

so that

$$v_0 = u_0 - x_0 = 2.0000 \tag{74}$$

As a reminder that the approximate evaluation of the superposition integral involves $v_0/2$, the first value of v in Table III is crossed out and replaced by half its value.

At $t = 0.20$ or $n = 1$,

$$Q_1 = (\alpha_1)\left(\frac{v_0}{2}\right) = (0.1637)(1.0000) = 0.1637 \tag{75}$$

$$\bar{Q}_1 = Q_1 + 0.2000 = 0.3637 \tag{76}$$

The equation

$$y_1 = \bar{Q}_1 - 0.10 x_1 = 0.3637 - 0.10 x_1 \tag{77}$$

is solved simultaneously with the diode characteristic curve, as shown in Fig. 9, giving

$$x_1 = 0.010 \tag{78}$$

$$y_1 = 0.362 \tag{79}$$

The corresponding value of v is

$$v_1 = u_1 - x_1 = 1.990 \tag{80}$$

At $t = 0.40$ or $n = 2$,

$$Q_2 = \alpha_1 v_1 + \alpha_2\left(\frac{v_0}{2}\right) = (0.1637)(1.990) + (0.1341)(1.000) = 0.4599 \tag{81}$$

$$\bar{Q}_2 = Q_2 + 0.2000 = 0.6599 \tag{82}$$

The equation

$$y_2 = 0.6599 - 0.10 x_2 \tag{83}$$

is solved simultaneously with the diode characteristic curve, as shown in Fig. 9, the values of x and y are entered in Table III, and the same process is repeated for successive values of n.

Similar calculations for $\Delta t = T = 0.10$ again give a more complete picture of the transient solution but no increase in accuracy, indicating that the graphical solution is again the limiting factor.

Linear Case

With a unit resistor substituted for the diode rectifier,

$$x_n = y_n \tag{84}$$

and equation 33 becomes

$$x_n = y_n = \frac{\bar{Q}_n}{1+\mu} \tag{85}$$

with $\mu = Tg_0/2$. Results obtained by using the proposed method, with equation 85 substituted for the graphical solution, are given in Table IV for comparison with values which have been computed from the exact solution

$$e_2(t) = 1 - e^{-1t} \tag{86}$$

obtained by ordinary operational methods.

Experimental Verification

Experimental results for both nonlinear and linear versions of this network are shown in Fig. 10. Agreement between computed and experimental values is again satisfactory.

Bridged-T Network

The concluding network example is the bridged-T network shown in Fig. 11(A). The element values used are $2R = 4{,}800$ ohms and $C = 1$ microfarad; in per-unit values, $2R = 2$ and $C = 1$. A 12-volt or 2-unit step will be applied to the input terminals of the network with the capacitors initially uncharged.

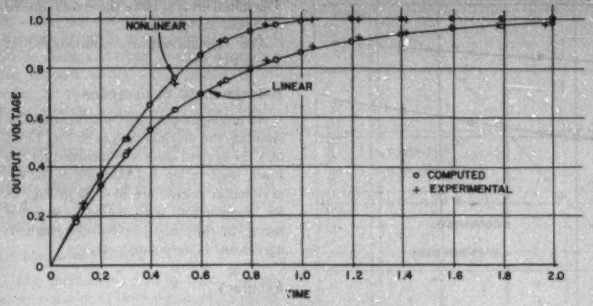


Fig. 10. Response of 1-section R-C ladder network

PRELIMINARY CALCULATIONS

In this case, the transfer function of the input block is

$$G_1(s)=1 \tag{87}$$

so that

$$u(t)=r(t)=e_1(t)=1.0000 \tag{88}$$

The feedback block is

$$H(s)=\frac{2s+1}{2s(s+1)} \tag{89}$$

so that

$$h(t)=\frac{1}{2}+\frac{1}{2}e^{-t} \tag{90}$$

For $\Delta t=T=0.20$, the β-sequence defined by equations 12 and 13 is

$$\frac{0.2000}{2},\ 0.1819,\ 0.1670,\ 0.1549,\ \ldots,\ 0.1000,\ \ldots \tag{91}$$

and $\rho=0.10$. The general solution for x_n is thus

$$x_n=S_n-0.10y_n \tag{92}$$

The initial conditions give

$$e_2(0)=x_0=y_0=1.0000 \tag{93}$$

After $y(t)$ has been determined, it is necessary to use another approximate evaluation of a superposition to find $a(t)$. The transfer function of the output block is

$$G_2(s)=\frac{-1}{s+1} \tag{94}$$

and the corresponding weighting function is

$$g_2(t)=-e^{-t} \tag{95}$$

For $\Delta t=T=0.20$, the necessary sequence of values is given by equation 69. After $a(t)$ is found, then $e_2(t)$ is found from

$$e_2(t)=c(t)=a(t)+b(t) \tag{96}$$

where

$$b(t)=r(t)=1.0000 \tag{97}$$

because $G_4(s)=1$.

TRANSIENT SOLUTION

The analysis of the feedback system for $y(t)$ follows the same pattern as in the first example and is not given in detail; the results are given in Table V. The calculation of $a(t)$ from $y(t)$ by the scheme of Fig. 2, using the weighting function of equation 95, is likewise straightforward. It will be noted that $y_6/2=0.500$ is used in both calculations.

LINEAR CASE

When a unit resistor is substituted for the nonlinear diode, a completely linear network is obtained which can be analyzed by the proposed method (using an algebraic solution instead of the graphical solution) and by ordinary operational methods. Values of e_2 obtained by both methods are given in Table VI. The approximate solution is remarkably good, even for $T=0.4$, considering that two approximate evaluations of superposition integrals were required.

EXPERIMENTAL VERIFICATION

Experimental results for both nonlinear and linear versions of this network are shown in Fig. 12. Agreement between computed and experimental results is acceptable, although not as good as in the previous examples.

Saturating Third-Order Servomechanism

The concluding example is a saturating third-order servomechanism. The block diagram, shown in Fig. 13(A), might represent a positioning system using a hydraulic motor with a limited pump stroke (y). Subject to some simplifying assumptions, the undamped resonant frequency ω_0 is dependent on the volume and bulk modulus of the oil, the inertia of the system, and the motor displacement; the damping ratio ζ depends on these same quantities and also on the leakage of the system.

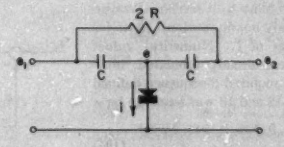
An equivalent block diagram for the special case $\omega_0=1.00$ and $\zeta=0.40$ is shown in Fig. 13(B); this diagram is a highly degenerate case of Fig. 1(A), with

$$G_1(s)=1 \tag{98}$$

$$H(s)=G_3(s)=\frac{1}{s(s^2+0.8s+1)} \tag{99}$$

$$G_2(s)=0 \tag{100}$$

The relation between the pump stroke y and the system error e will be assumed to be



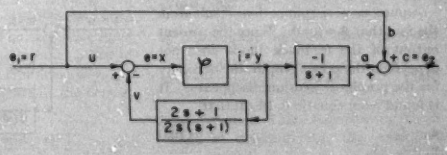


Fig. 11. Bridged-T network and block diagram

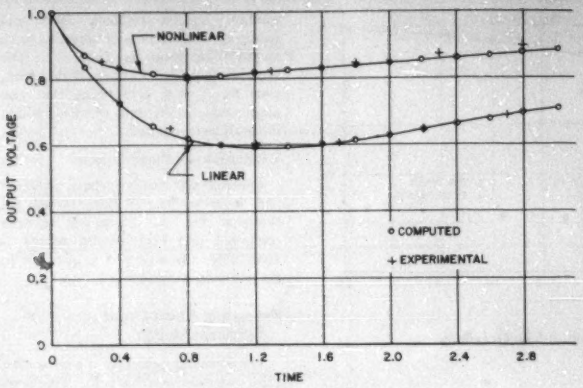


Fig. 12. Response of bridged-T network

$$y = \begin{cases} +Ke_c & e > e_c \\ Ke & |e| \leq e_c \\ -Ke_c & e < -e_c \end{cases} \quad (101)$$

with $K = 0.5$ and $e_c = 2$; y therefore varies between plus and minus one. When unsaturated, the system operates as a linear system with a transfer function

$$\frac{C(s)}{R(s)} = \frac{0.5}{s^3 + 0.8s^2 + s + 0.5} \quad (102)$$

Preliminary Calculations

The weighting function for the feedback block, the inverse transform of equation 99, is both the impulse response of a third-order system and the step response of a second-order system. As a result, the inverse transform

$$h(t) = 1 - e^{-\zeta\omega t}\left[\cos\sqrt{1-\zeta^2}\,\omega t + \frac{\zeta}{\sqrt{1-\zeta^2}}\sin\sqrt{1-\zeta^2}\,\omega t\right] \quad (103)$$

is available in almost all servomechanism texts[38] in handy nondimensional plots for various values of ζ. Numerical values for $\zeta = 0.4$ were already tabulated, however, and the required β-sequence defined by equations 12 and 13 was found to be

$$0.0000,\ 0.0284,\ 0.0992,\ 0.1916,\ 0.2864,\ \ldots \quad (104)$$

for $\Delta t = T = 0.40$.

A number of simplifications result from the fact that $\beta_0 = \rho = 0$. Since the present output of the feedback block does not depend on the present input, the necessity for the graphical solution disappears. It is found from equations 16–20 that

$$u_n = S_n \quad (105)$$

$$x_n = u_n - S_n = S_n \quad (106)$$

These may be traced to the fact that

$$\lim_{s \to \infty} [sH(s)] = 0 \quad (107)$$

a condition which did not exist in the other examples.

In addition, because of the nature of the original system, it is also true that

$$c(t) = v(t) \quad (108)$$

eliminating the need for a separate analysis to obtain $c(t)$.

Calculations

The calculations, details of which are not given, were carried out for a 4-unit step function applied to the input with the system initially at rest, so that

$$r(t) = u(t) = 4.0000 \quad (109)$$

$$e(0) = x_0 = 4.0000 \quad (110)$$

$$y_0 = 1.0000 \quad (111)$$

The initial error is twice the error required to cause saturation.

The calculations are summarized in Table VII.

Experimental Verification

An analogue computer was arranged to simulate the original system of Fig. 13(A) and the system error was recorded for the input used in the calculations. The experimental results are plotted in Fig. 14 for comparison with values computed using the proposed step-by-step method; agreement is very satisfactory.

Accuracy

Discrepancies between computed values and the corresponding experimental data in work on this kind can be traced to a number of sources and can be classified as (1) experimental errors, (2) representation errors, and (3) analytical errors.

Experimental errors include such things as errors in the values of the fixed linear elements or in the characteristic curves of the nonlinear elements, errors due to static or dynamic limitations of the recording equipment, and errors in transcribing the experimental data.

Representation error is a term used to characterize errors resulting from failure of the mathematics to describe completely the physical system or conversely. In the circuit problems, for example, the capacitance of the rectifiers was neglected; its effect could be included, at least approximately, by starting with a more complex equivalent circuit for the rectifier. In the servomechanism example, on the other hand, the analogue computer was supposed to simulate the system assumed in the analysis but could not do so with complete accuracy because of operational amplifier limitations, potentiometer loading, and so forth.

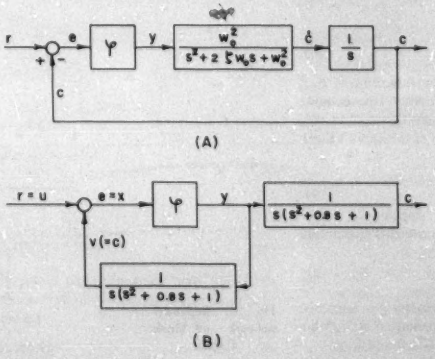


Fig. 13. Block diagrams for saturating third-order servomechanism

Analytical errors include mistakes in the arithmetic and graphical work, round-off errors, errors in the numerical integration formula, and errors due to accumulation of past errors. The possibility of errors of the first three types can be avoided by elementary stratagems: difference checks of the variables as they are obtained, use of a sharp pencil and large sheets of graph paper, and carrying more figures than will be retained in the final results. The last two types of error deserve a more careful examination.

Integration Errors

The trapezoidal rule is based on an approximation of the integrand by a series of straight-line segments and will therefore give correct results for any integrand which has a constant first derivative. Any errors which arise can thus be traced to the higher derivatives of the integrand A thorough investigation of this matter leads to the following estimate of the error, defined as the observed value minus the true value:[20]

$$E_1 \leqslant +\frac{LT^2}{12} Z_m'' \tag{112}$$

where

T = interval between values of independent variable
$L = nT$ = total interval of integration
Z = function being integrated (integrand)
Z_m = maximum value of the second derivative of the integrand

A more optimistic estimate of the error is[19] the equation

$$E_2 \leqslant +\frac{T^2}{12}\{Z'(nT) - Z'(0)\} \tag{113}$$

where $Z'(0)$ and $Z'(nT)$ are the first derivatives of the integrand at the beginning and end of the interval. Since the first derivative is the integral of the second,

$$Z'(nh) - Z'(0) = \int_0^{nT} Z'' dt \tag{114}$$

$$= LZ_{av}'' \tag{115}$$

where Z_{av}'' denotes the average value of the second derivative of the integrand in the interval under consideration. Equation 113 may therefore be written in the alternative form

$$E_2 \leqslant +\frac{LT^2}{12} Z_{av}'' \tag{116}$$

The positive sign preceding equations 112 and 116 reflects the fact that the approximate value of the integral will be too large if the curve representing the integrand is concave upward (positive second derivative), placing the straight-line approximation above the actual curve.

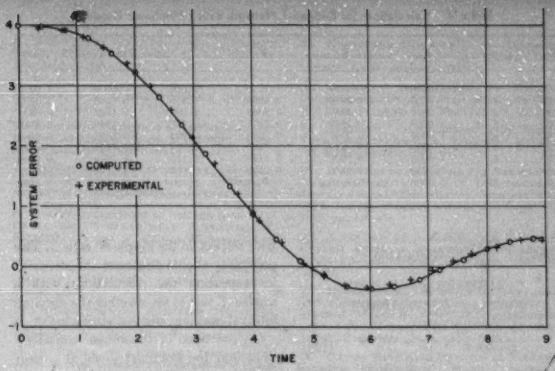


Fig. 14. Response of saturating third-order servomechanism

The estimate derived from equation 112 is unduly pessimistic and is, in addition, more difficult to apply since it requires determination of the maximum value of the second derivative instead of the difference of two easily located first derivatives. Equation 113 on the other hand gives a fairly accurate estimate of the actual error, particularly for reasonable values of T, and is much easier to use.

For the present application, the integral to be evaluated is either

$$v(t) = \int_0^t h(\tau) y(t-\tau) d\tau \tag{117}$$

or

$$y(t) = \int_0^t g(\tau) v(t-\tau) d\tau \tag{118}$$

depending on the form of the block diagram. If, in equation 117, one sets

$$Z(\tau) = h(\tau) y(t-\tau) \tag{119}$$

then the first derivative (with respect to τ) is

$$Z'(\tau) = -h(\tau) y'(t-\tau) + h'(\tau) y(t-\tau) \tag{120}$$

Letting $t = nT$ and substituting equation 120 into equation 113, obtains

$$E\{v(nT)\} \leqslant \frac{T^2}{12}[h(0)y'(nT) - h'(0)y(nT) - h(nT)y'(0) + h'(nT)y(0)] \tag{121}$$

as an estimate of the error.

While $h(t)$ and its derivative are known in advance, $y(t)$ is determined in the course of the calculations and its derivative must be estimated from the calculated values of $y(t)$, either graphically or numerically. A fairly satisfactory estimate of the derivative can be obtained from

$$y'(nT) \cong \frac{y[(n+1)T] - y[(n-1)T]}{2T} \tag{122}$$

if the necessary values are available.

A similar set of relations applicable to equation 118 can of course be derived, involving $g(t)$, $v(t)$, and their derivatives.

Example

To illustrate the application of these relations, the error in the output voltage of the 2-section R-C ladder network will be estimated. A linear case is used in order to have an exact result and the actual error available for comparison.

In this network, the weighting function was

$$h(t) = 2e^{-t} \tag{123}$$

with the first derivative

$$h'(t) = -2e^{-t} \tag{124}$$

If an estimate of the error in $v(1)$ is desired, then (by direct evaluation)

$$h(0) = 2.0000 \qquad h(1) = 0.7358$$
$$h'(0) = -2.0000 \qquad h'(1) = -0.7358 \tag{125}$$

and (by reference to an appropriate table)

$$y(0) = 0.0000$$
$$y(1) = 0.9494 \tag{126}$$

The necessary derivatives of $y(t)$ are estimated to be

$$y'(0) = 0.350 \tag{127}$$

by plotting the first few values of y, and

$$y'(1) = \frac{0.9713 - 0.9092}{(2)(0.2)} = 0.155 \tag{128}$$

from equation 122. Substituting these values in equation 121, the estimated error in $v(1)$ is

Table V. Calculations for Bridged-T Network with a Nonlinear Diode

t	n	u	S	$\bar{S}$	y	s	e_T
0	0	1.0000			0.500	0.0000	1.0000
0.2	1	1.0000	0.0910	0.9090	0.450	0.1248	0.8752
0.4	2	1.0000	1.1617	0.8383	0.295	0.1669	0.8331
0.6	3	1.0000	0.2029	0.7971	0.233	0.1842	0.8158
0.8	4	1.0000	0.2307	0.7693	0.200	0.1899	0.8101
1.0	5	1.0000	0.2517	0.7483	0.170	0.1888	0.8112
1.2	6	1.0000	0.2670	0.7330	0.155	0.1840	0.8160
1.4	7	1.0000	0.2800	0.7200	0.143	0.1776	0.8224
1.6	8	1.0000	0.2912	0.7088	0.130	0.1701	0.8299

$$E\{v(1)\} \lesssim \frac{(0.2)^2}{13}\,[(2.0000)(0.155) - (-2.0000)(0.9494) - (0.7358)\times(0.350) + (-0.7358)(0.0000)] \lesssim (0.0033)(1.950) \lesssim 0.0064 \qquad (129)$$

Since the output voltage is one half of $v(t)$, the estimated error in $e_2(1)$ is therefore

$$E\{e_2(1)\} \lesssim 0.0032 \qquad (130)$$

The actual error, according to Table II, is 0.0004, appreciably less than the estimate given by equation 130.

The error estimates from equation 121 are useful only if the integration formula is the chief source of error. In linear systems where errors in the graphical solution can be eliminated and an exact value obtained for comparison, as in Tables II, IV, and VI, the estimates have proved consistently conservative. Moreover, inspection of these tables reveals that the error in a particular value does vary as the square of the interval T, giving support to the belief that the integration formula is the chief source of error in this method.

In nonlinear problems, the main source of error is the graphical solution. With an error estimate based on a preliminary set of calculations and knowing that the integration error varies as T^2, a new interval can be chosen (if necessary) which will render the integration error negligible with respect to the graphical error. This possibility should eliminate the necessity for repeating the calculations with a smaller T just to see whether the previous interval was already satisfactory.

On the basis of theoretical considerations and the examples given, it is clear that the approximate solutions become exact as the interval T approaches zero. On the other hand, sample calculations show that the results can become completely useless if T is made excessively large. The error estimates from equation 121 are useful only if the values are reasonably good to start with; this can usually be determined by inspection.

Selection of a suitable value of T must evidently be based on $h(t)$ and on the rate of variation of y with time, which cannot be predicted in advance because it depends on the nature of the nonlinearity, the input, and the linear parameters. The examples given and further experience with the method will provide a guide to a wise choice of T for the first set of calculations.

Table VI. Results of Calculations for Bridged-T Network with a Linear Resistor

t	Output Voltage: Approximate T = 0.20	Approximate T = 0.40	Exact
0	1.0000	1.0000	1.0000
0.2	0.8356		0.8358
0.4	0.7280	0.7271	0.7284
0.6	0.6605		0.6609
0.8	0.6208	0.6202	0.6212
1.0	0.6005		0.6008
1.2	0.5935	0.5932	0.5937
1.4	0.5935		0.5959
1.6	0.6036	0.6035	0.6036
1.8	0.6155		0.6155
2.0	0.6299	0.6299	0.6298

Accumulated Errors

With the main source of error already identified, it may seem unnecessary to look for additional sources. Nevertheless, it is worth while to examine the question: Does an error, once made, lead to decreasing or increasing errors in the following values of the variables?

In an attempt to answer this question, deliberate errors were introduced into early values in tables such as those of Tables I, III, or V, and the subsequent calculations repeated. For the cases examined, the effects of these deliberate errors eventually disappeared. Further investigation is required to determine the conditions under which this self-correction of errors is obtained.

Conclusions

A general method has been described for the transient analysis of systems having one input, one output, and one nonlinear element characterized by a relation between the instantaneous values of its input and output, all of which can be represented by the block diagrams of Fig. 1. The block diagram formulation of the problem is used not because it leads to any unique results, but because it seems to facilitate a standardized attack on problems in the analysis of nonlinear feedback control systems, electric networks, and a variety of analogous dynamic systems.

The block diagram approach has some advantages, however, since it may suggest a formulation of a problem which is more convenient than the conventional formulation. In particular, the transformation rule which permits the interchange of the forward and feedback blocks insures that the weighting functions will be free of troublesome impulsive components.

The analytical method itself uses only operational mathematics, numerical integration, and graphical solution of algebraic equations, techniques generally familiar to engineers. The method thereby gains in convenience for occasional use but probably suffers in efficiency. The amount of labor required by the proposed method compares rather unfavorably with other methods for which the number of operations per step is a constant. In defense of the method, it may be noted that: (1) the number of operations is independent of the order of the system, although the preliminary calculations become more involved as the system complexity increases; (2) the preliminary calculations can make use of existing tables or curves, and the weighting function sequences can be saved for re-use; (3) if the weighing function approaches zero or a constant for large values of t, the weighted sum of sufficiently distant values of y also approaches zero or a constant, and the number of operations per step can become a constant. Since the preliminary

Table VII. Calculations for Saturated Third-Order Servomechanism

t	u=r	S=v=c	$\bar{S}$=x=e	y
0	4.0000	0.0000	4.0000	0.5000
0.4	4.0000	0.0142	3.9858	1.0000
0.8	4.0000	0.0780	3.9220	1.0000
1.2	4.0000	0.2234	3.7766	1.0000
1.6	4.0000	0.4624	3.5376	1.0000
2.0	4.0000	0.7910	3.2090	1.0000
2.4	4.0000	1.1944	2.8056	1.0000
2.8	4.0000	1.6516	2.3484	1.0000
3.2	4.0000	2.1402	1.8598	0.9299
3.6	4.0000	2.6660	1.3340	0.6670
4.0	4.0000	3.1170	0.8830	0.4418
4.4	4.0000	3.5483	0.4517	0.2258
4.8	4.0000	3.9039	0.0961	0.0480
5.2	4.0000	4.1634	−0.1634	−0.0817
5.6	4.0000	4.3171	−0.3171	−0.1586
6.0	4.0000	4.3670	−0.3670	−0.1835

calculations are independent of the nonlinear element, these calculations do not have to be repeated for each new nonlinearity; it is relatively easy to examine the effects of a variety of nonlinear functions (which need not be expressed analytically) in the same system.

Evaluation of the superposition integral by numerical integration techniques naturally suggests use of standard error estimation formulas, making it possible to judge the accuracy of the results and to select appropriate step intervals. This desirable goal is achieved with greater difficulty, if at all, by other similar methods.[25–27,29,37]

Examples have been given illustrating variations of the method which may be encountered if $y(0)$ and $h(0)$ are or are not zero and if some of the transfer functions become zero, unity, or identical with other transfer functions. The results of the calculations check satisfactorily with experimental data.

The method can obviously be extended to systems in which $y = \varphi(x)$ becomes $y = \varphi'(x, t)$, simply by substituting a family of curves (with time as the parameter) for the single curve used in these examples. Time-varying linear and nonlinear elements can both be treated in this way. It is also possible that an adaptation of the method may be applicable to some nonlinear sampled-data systems.

References

1. Ordinary Nonlinear Differential Equations in Engineering and Physical Sciences (book), N. W. McLachlan. Clarendon Press, Oxford, England, 1950.

2. D-C Transient Response of Single Energy Nonlinear Circuits, E. Weber. *Research Report R-271-52, PIB-210*, Microwave Research Institute, Polytechnic Institute of Brooklyn, Brooklyn, N. Y., July 28, 1952.

3. Differentialgleichungen: Losungsmethoden und Losungen (book), E. Kamke. Chelsea Publishing Company, New York, N. Y., 1948.

4. Introduction to Non-Linear Mechanics (book), Nicholas Minorsky. J. W. Edwards, Ann Arbor, Mich., 1947.

5. Theory of Oscillations (book), A. A. Andronow, C. E. Chaikin. Princeton University Press, Princeton, N. J., 1949.

6. Nonlinear Vibrations in Mechanical and Electrical Systems (book), J. J. Stoker. Interscience Publishers, New York, N. Y., 1950.

7. Erzwungene Schwingungen bei veranderlich Eigenfrequenz (book), G. Duffing. F. Vieweg und Sohn, Braunschweig, Germany, 1918.

8. Steady Oscillations of Systems with Nonlinear and Unsymmetrical Elasticity, M. Rauscher. *Journal of Applied Mechanics*, New York, N. Y., vol. 5, 1938.

9. Les Methodes Nouvelles de la Mecanique Celeste, J. H. Poincare. Gauthier-Villars, Paris, France, 1892.

10. Introduction to Nonlinear Mechanics, N. Kryloff, N. Bogoliuboff. Translation by S. Lefschetz. Princeton University Press, 1943.

11. A Method of Equivalent Linearization for Nonlinear Oscillatory Systems with Large Nonlinearity, C. A. Ludeke. *Journal of Applied Physics*, New York, N. Y., vol. 20, July 1949, pp. 694–99.

12. A Frequency Response Method for Analyzing and Synthesizing Contactor Servomechanisms, Ralph J. Kochenburger. *AIEE Transactions*, vol. 69, 1950, pp. 270–84.

13. Operational Methods in Nonlinear Mechanics, L. A. Pipes. *Report 51–10*, Department of Engineering, University of California, Los Angeles, Calif., Dec. 1951.

14. The Reversion Method for Solving Nonlinear Differential Equations, L. A. Pipes. *Journal of Applied Physics*, New York, N. Y., vol. 23, Feb. 1952, pp. 202–07.

15. A Mathematical Analysis of a Series Circuit Containing a Nonlinear Capacitor, Louis A. Pipes. *AIEE Transactions*, vol. 72, pt. I, July 1953, pp. 238–44.

16. Mathematics of Modern Engineering—Vol. II (book), E. G. Keller. John Wiley & Sons, Inc., New York, N. Y., 1942.

17. Theory and Calculation of Variable Electrical Systems, J. R. Carson. *Physical Review*, New York, N. Y., vol. 17, Feb. 1921, pp. 116–34.

18. Applications of Integral Equations to the Solution of Nonlinear Electric Circuit Problems, L. A. Pipes. *AIEE Transactions*, vol. 72, 1953, pp. 445–50.

19. Numerical Mathematical Analysis (book), J. B. Scarborough. Johns Hopkins Press, Baltimore, Md., first edition, 1930.

20. Numerical Calculus (book), W. E. Milne. Princeton University Press, 1949.

21. Numerical Solution of Differential Equations (book), W. E. Milne. John Wiley & Sons, Inc., 1953.

22. Numerische Behandlung von Differentialgleichungen (book), L. Collatz. Julius Springer, Berlin, Germany, 1951.

23. Transient Performance of Electric Power Systems (book), R. Rudenberg. McGraw-Hill Book Company, Inc., New York, N. Y., 1950.

24. Graphical Constructions for Vacuum Tube Circuits (book), A. Preisman. McGraw-Hill Book Company, Inc., 1943.

25. A Method of Analyzing the Behavior of Linear Systems in Terms of Time Series, A. Tustin. *Journal*, Institution of Electrical Engineers, London, England, vol. 94, no. 1, pt. IIA, 1947, pp. 130–42.

26. A Method of Analyzing the Effect of Certain Kinds of Nonlinearity in Closed-Cycle Control Systems, A. Tustin. *Ibid.*, pp. 152–60.

27. Number Series Method of Solving Linear and Nonlinear Differential Equations, A. Madwed. *Report No. 6443-T-26*, Instrumentation Laboratory, Massachusetts Institute of Technology, Cambridge, Mass., Apr. 1950.

28. On the Numerical Solution of Volterra Integral Equations, C. Wagner. *Journal of Mathematics and Physics*, Massachusetts Institute of Technology, Cambridge, Mass., vol. 32, Jan. 1954, pp. 289–301.

29. Numerical Analysis for Network Design, J. G. Truxal. *PGCT Transactions*, Professional Group on Circuit Theory, Institute of Radio Engineers, New York, N. Y., vol. CT-1, Sept. 1954, pp. 49–60.

30. Block Diagram Transformations for Systems with One Nonlinear Element, T. M. Stout. *AIEE Transactions*, vol. 75, pt. II, July 1956, pp. 130–41.

31. Transients in Linear Systems (book), M. F. Gardner, J. L. Barnes. John Wiley & Sons, Inc., 1942.

32. Open-Loop Frequency Response Method for Nonlinear Servomechanisms, R. L. Cosgriff. *AIEE Transactions*, vol. 72, pt. II, Sept. 1953, pp. 222–35.

33. A Generalized Method for Determining the Closed-Loop Frequency Response of Nonlinear Systems, Luther R. Prince, Jr. *Ibid.*, vol. 73, pt. II, Sept. 1954, pp. 217–24.

34. Sinusoidal Analysis of Feedback-Control Systems Containing Nonlinear Elements, E. Calvin Johnson. *Ibid.*, vol. 71, pt. II, July 1952, pp. 169–81.

35. A Method of Approximate Steady-State Analysis of Nonlinear Networks, G. R. Sisson. *Proceedings*, Institution of Electrical Engineers, London, England, vol. 100, pt. I, Sept. 1953, pp. 275–87.

36. Principles of Servomechanisms (book), G. S. Brown, D. P. Campbell. John Wiley & Sons, Inc., 1948.

37. Automatic and Manual Control (book), B. M. Brown. "Application of Finite Difference Operators to Linear Systems." Academic Press, New York, N. Y., 1952, pp. 409–19.

Discussion

G. W. Bills (North American Aviation, Inc., Downey, Calif): This method for the transient analysis of feedback systems containing one nonlinear element has the advantage of simplicity when it is applied to a problem. It appears to be more suited for calculation by ordinary means using such aids as an ordinary desk calculator than by attempting to program it for an International Business Machines Corporation *704*.

On one voltage regulator problem which involved response times under 20 milliseconds, we tried T intervals as low as 0.25 millisecond and still experienced erratic output values. Apparently some discretion is necessary when applying this method to a system to be certain that the time constants involved are not too small. When the Thaler and Boxer method was used for the same problem the employment of a time interval of 1 millisecond gave consistent responses.

Although it was stated in this paper that "the present method is applicable to systems having one input," there seems to be no reason why it cannot be extended to multiple inputs. We experienced little difficulty in applying it to a 2-input system other than that previously mentioned with regard to the time interval.

For those nonlinear problems where a suitable and practical time interval can be chosen, this method appears to offer an easy way of obtaining solutions.

T. M. Stout: This discussion by G. W. Bills is very much appreciated. It is particularly gratifying to find that he and his associates have applied this method to some of their problems. Without additional information concerning his voltage regulator problem, it is difficult to guess why the interval $T = 0.25$ millisecond did not produce satisfactory results, and this reply therefore consists of some general remarks which may relate to his comments.

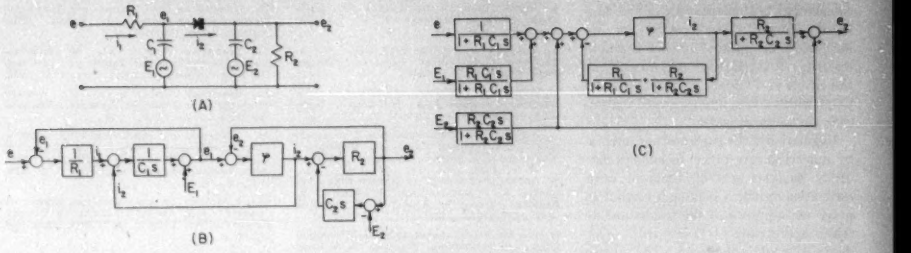


Fig. 15

A—Network of Fig. 5 showing voltages E_1 and E_2 which can represent initial capacitor voltages or additional inputs
B—Original block diagram
C—Final block diagram

In a paper[1] which appeared after this work was done, Boxer and Thaler use the substitution

$$\frac{1}{s}=\frac{T}{2}\left[\frac{1+z^{-1}}{1-z^{-1}}\right] \tag{131}$$

and more complicated expressions for s^{-2}, s^{-3}, . . .to convert Laplace transforms into z-transforms or difference equations. If, for example,

$$u=\frac{1}{s}\,y \tag{132}$$

use of equation 131 gives

$$u=\frac{T}{2}\left[\frac{1+z^{-1}}{1-z^{-1}}\right]\,y \tag{133}$$

or

$$u_n=u_{n-1}+\frac{T}{2}[y_n+y_{n-1}] \tag{134}$$

which is, as Boxer and Thaler point out, a form of the trapezoidal rule. (The relations between z-transforms and the numerical integration formulas are also discussed by Salzer.[2]) In developing the corresponding substitutions for multiple integrations Boxer and Thaler do not (as Tustin[3] does) use powers of equation 131 but instead develop new relations which do not appear to correspond to any of the better-known numerical integration formulas. Results obtained by their method are probably more accurate than can be obtained using only the trapezoidal rule, and it is likely that a larger interval could be used for the same accuracy. It is not clear whether the procedure described in the paper for estimating errors could be applied if the actual numerical work is performed using exclusively powers of equation 131 and the resulting difference relations, or whether the ability to estimate the errors would be adequate compensation for the probable loss in accuracy.

Iustead of the trapezoidal rule,

$$\int_0^{nT} z\,dr=\frac{T}{2}(z_0+2z_1+2z_2+\ldots+2z_{n-1}+z_n)$$

it would be possible to use a combination of Simpson's one-third rule and the trapezoidal rule, giving

$$\int_0^{nT} z\,dr=\frac{T}{6}(3z_0+3z_1) \qquad n=1$$

$$=\frac{T}{6}(2z_0+8z_1+2z_2) \qquad n=2$$

$$=\frac{T}{6}(2z_0+8z_1+5z_2+3z_3) \qquad n=3$$

$$=\frac{T}{6}(2z_0+8z_1+4z_2+8z_3+2z_4) \qquad n=4$$

and so on. This procedure gives increased accuracy in exchange for greater complexity; several different strips would be required instead of the single strip shown in Fig. 2. Since a larger interval could be used for the same accuracy, the total labor might not be excessive.

The method described in this paper differs from that of Boxer and Thaler in several other respects. Their method, like Madwed's,[4] starts generally with a differential equation, while the method described here resembles Tustin's[5] in starting with a block or signal-flow diagram. They consider nonlinear systems in terms of differential equations with varying coefficients, while here the nonlinearity is represented by a transfer relation and the differential equations do not appear explicitly. They advocate an iterative procedure to avoid errors due to use of y_{n-1} in place of the unknown present value y_n; this problem is avoided in the present method by use of the graphical construction. The methods also differ in respect to the procedures for dealing with initial conditions.

In this method, initial conditions can be treated as additional inputs. For example, a charged capacitor can be treated as an uncharged capacitor in series with a battery, and similar tricks can be applied to other types of element.[6] An example, based on the 2-section R-C ladder network, is given in Fig. 15. The voltages E_1 and E_2, which may represent either initial capacitor voltages or additional voltage sources, can be included in the block diagram for the net work by noting that

$$e_1=E_1+\frac{1}{C_1}\int(i_1-i_2)dt \tag{135}$$

$$i_{C2}=C_2\frac{d}{dt}(e_2-E_2) \tag{136}$$

After a number of steps (not shown), the original block diagram of Fig. 15(B) can be reduced to the form given in Fig. 15(C). The voltage E_2 contributes to the input to the nonlinearity and to the final output, while E_1 contributes only to the input to the nonlinearity. The problem of obtaining the output of the nonlinearity for a specified e, E_1, and E_2 is no more difficult than before, and the same general procedure can be used.

Reflection on the general diagram of Figs. 1(A) and 1(B) (developed in a previous paper[7]) indicates that each additional input will, in general, produce another signal, like u, at the input of the nonlinearity and a signal, like b, at the final output of the system. The transfer functions involved will be, in general, different and bear the same relationship to the additional signal that G_1 and G_2 bear to r.

I am inclined to agree that the proposed method is better adapted to hand computation than to high-speed digital computers. It should be suitable for occasional use when a standard, easily remembered method is desired.

References

1. A Simplified Method of Solving Linear and Nonlinear Systems, R. Boxer, S. Thaler. *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 44, Jan. 1956, pp. 89–101.
2. Frequency Analysis of Digital Computers Operating in Real Time, J. M. Salzer. *Ibid.*, vol. 42, Feb. 1954, pp. 457–66.
3. Reference 25 of the paper.
4. Reference 27 of the paper.
5. Reference 26 of the paper.
6. Initial Conditions in Transient Analysis, L. Tasny-Tschiassny. *Wireless Engineer*, London, England, vol. 30, Feb. 1953, pp. 28–31.
7. Reference 30 of the paper.

Hidden Oscillations in Sampled-Data Control Systems

ELIAHU I. JURY
ASSOCIATE MEMBER AIEE

Synopsis: Oscillations between sampling instants in the response of sampled-data control systems impose certain limitations on the z-transform method and other methods available for synthesis of such systems. It is shown that these oscillations appear in the response because of certain forms of the open-loop transfer function $G(z)$. If the impulsive response of these functions is oscillatory then these oscillations occur in the response, and if the sampling period is larger or equal to half the period of these oscillations, the z-transform method fails to indicate their presence. The modified z-transform can be advantageously applied in these cases. This method is straightforward and simple in detecting the presence of these oscillations and indicating their values. Theorems relating to initial and final values of the response in the form of modified z-transform has been presented. Two examples have been discussed in which such oscillations occur and the modified z-transform method has been applied in their analysis.

IT IS known that the z-transform method can be applied systematically for analysis and synthesis of sampled-data control systems. However, the design based entirely on the z-transform method does not yield any information between sampling instants. This fact has led a few authors[1-6] to introduce new methods to obtain either exact or approximate values of the response between the sampling instants. The very promising method developed is called the modified z-transform[2] and can be applied advantageously for sampled-data synthesis and analysis. The problem of when the z-transform method can be simply applied depends mainly on the behavior of the systems between the sampling instants. If the response is well behaved between the sampling instants then the z-transform can be directly applied. However, if the response between the sampling instants is highly oscillatory the z-transform can be very misleading as shown in this paper.

The purpose of this paper is to discuss in detail the systems behavior between the sampling instants, and to devise methods of detection which give notice of these hidden oscillations between sampling instants and to determine whether the z-transform method or other methods can be applied.

The main investigation is based on the impulsive form of $G(s)$, the open-loop transfer function of the control system. Based on this approach, a method of detecting hidden oscillations has been developed.

Once the hidden oscillations are known to exist, then one can apply directly the modified z-transform to measure the magnitude and occurrences of these undesirable oscillations.

A method has been indicated in which the z-transform can be applied under certain conditions although these oscillations exist.

Thus, the object of this paper is mainly to clarify the choice between the z-transform or the modified z-transform to be applied in the analysis and synthesis of sampled-data control systems.

Oscillations between Sampling Instants

The choice between the use of the z-transform or the modified z-transform method for analysis and synthesis of sampled-data systems mainly depends upon the systems response between sampling instants. In most cases, the response between sampling instants follow the same pattern of the response at the sampling instants, then the z-transform can be readily applied. However, there are cases in which the response is oscillatory between the sampling instants, then the z-transform method fails to indicate these oscillations and the modified z-transform is best applicable under these conditions.

The occurrences of these high oscillations between the sampling instants where the z-transform method fails to indicate, can be mainly attributed to the following factors:

1. The value of the sampling period T.
2. The form of $G(s)$, the open-loop transfer function.

In most sampled-data feedback systems (configuration of which is indicated in Fig. 1), the form of $G(s)$ is such that its impulsive response in general can have one of the following four shapes.

1. The impulsive response decays exponentially and this occurs when the degree of the numerator in s is one less than the degree of the denominator. This assumes the roots of the denominator are real.
2. The impulsive response rises exponentially and this occurs when the degree of the denominator in s is at least two degrees higher than the numerator, as is encountered in most servo systems, and this assumes the roots of the denominator are real.
3. The impulsive response is oscillatory and this occurs when the poles $G(s)$ are complex. It is to be noticed that response of this form is for the general system and is generally valid if dominance holds.[7]
4. The impulsive response is oscillatory but the highest overshoot occurs not at the first maximum. This case is similar to the previous one except that dominance[7] does not hold.

It is to be noticed that only the last two cases can produce hidden oscillations, and then only if the period of oscillation is less than twice the sampling period.

Figs. 2–5 indicate the various impulsive responses of $G(s)$ where it is to be noticed the last two figures indicate hidden oscillations.

Furthermore, it is to be observed that the impulsive response in all these cases is bounded and consequently no instability occurs in the open-loop response but only oscillations.

Although the system is in feedback form, for the first sampling period the system behaves identically like an open-loop system from which fact one can readily deduce that if the sampling period is larger than one half of the first period of oscillation, the pulsed output will not indicate these oscillations and therefore the z-transform method can be misleading.

Furthermore, if one examines both $G^*(s)$ and $G^*(z,m)$ in these cases, it will be noticed that the poles of these functions may not be identical or equal in number. This indication in most cases

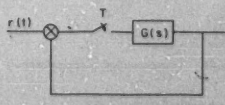


Fig. 1. Typical sampled-data control system

Paper 56-790, recommended by the AIEE Feedback Control Systems Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Summer and Pacific General Meeting, San Francisco, Calif., June 25–29, 1956. Manuscript submitted January 27, 1956; made available for printing April 27, 1956.

ELIAHU I. JURY is with the University of California, Berkeley, Calif.

This research was supported by the United States Air Force, through the Air Force Office of Scientific Research, Air Research and Development Command under Contract *AF 18(600)-1521*.

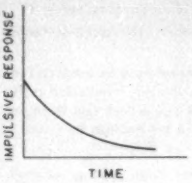


Fig. 2 (left). Monotonic impulsive response

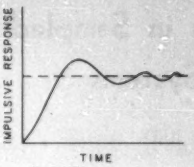


Fig. 4 (right). Nonmonotonic impulsive response

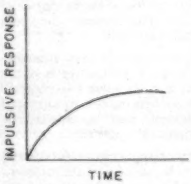


Fig. 3 (left). Monotonic impulsive response

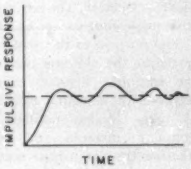


Fig. 5 (right). Nonmonotonic impulsive response

serves as a notice of occurrences of such oscillations. This is illustrated in Table I where the z-transform as well as the modified z-transform is indicated for various forms of $G(s)$ and is further clarified in Appendix I.

Hidden Instability or Hidden Oscillations?

In the literature,[1] these cases of severe oscillations between sampling instants have been inadvertently referred to as hidden instability; i.e., the sampled-data system can be unstable between sampling instants, although stable at the sampling time.

In examining closely the cause for this instability, one may readily deduce from the previous considerations that the impulsive response of the system is unstable. This fact is an indication that the sampled-data system is evidently unstable. Therefore, in these cases one should examine the impulsive response of $G(s)$ before applying the z-transform. Further, if $\mathcal{L}^{-1}\,G(s)$ is oscillatory but stable, the z-transform will not indicate these oscillations under the conditions mentioned earlier, hence the term, hidden oscillations, is introduced.

In the case of sampled-data control systems as shown in Fig. 1, one may deduce similar observations as follows:

1. If $G(s)$ is stable, then $G^*(z)$ which can be written as ratio of two polynomials in z

$$G^*(z)=\frac{A(z)}{B(z)} \tag{1}$$

is also stable. The modified z-transform of $G(s)$ is

$$G^*(z,m)=\frac{A(z,m)}{B_1(z)} \tag{2}$$

where $B_1(z)$ may have more roots than $B(z)$ depending on the impulsive response of $G(s)$. The over-all transfer function is

$$\frac{G^*(z,m)}{1+G^*(z)}=\frac{A(z,m)}{B(z)+A(z)}\times\frac{B(z)}{B_1(z)} \tag{3}$$

From equation 3, one may deduce the following: If the roots of $B_1(z)$ are within the unit circle, the stability of the system depends on the roots of $B(z)+A(z)$. Thus, if the impulsive response of $G(s)$ is stable no hidden instability exists. (The response between sampling instant cannot be unstable if it is stable at the sampling instants.)

2. Consider the case where $B_1(z)$ has roots outside the unit circle which do not appear in $B(z)$, then the response is unstable between sampling instants and can be stable at the sampling instants. But in examining this case, it is known *a priori* that the impulsive response of $G(s)$ is unstable if $B_1(z)$ has roots larger than unity, and consequently the modified z-transform is unstable. Conditions 1 and 2 are illustrated in example in Appendix I.

Oscillatory Form of g(t)

The conditions underlying the oscillatory form of $\mathcal{L}^{-1}\,G(s)$ can be attributed to the following:

1. The given form of $G(s)$ consists of complex poles.
2. Given the form of $G(s)$ having real poles, but the sampled data configurations contains multiple loops as shown in Fig. 6. In this case the open-loop transfer function of this system is $G(s)/1+G(s)$, which introduces complex zeros of $1+G(s)$ if the gain is sufficiently high.

The sampled-data configuration shown in Fig. 6 has not been discussed in detail in the literature.[8] Hence the preceding discussion will serve to illustrate that such systems will have oscillation between the sampling periods, and the z-transform method cannot be simply applied although $G(s)$ has nonoscillatory impulsive response.

The conditions for monotonic and nonmonotonic responses have been discussed in detail in reference 9 which gives conditions on these oscillations.

The Effect of the Sampling Period "T"

Usually the sampling period is fixed by the frequency spectrum of the input, which requires that T should be less or equal[10] to $1/2\omega$, where ω is the highest frequency of the input. However, if the sampling period is not so fixed, then one can choose its value to be equal or less than half the period of oscillations of the system impulsive response.

Thus, by such choice of T one can still apply the z-transform method advantageously although it is known that severe oscillations exist. This fact is of importance mainly in the analysis of linear or nonlinear continuous systems where the choice of T can be changed at will.

Application of the Modified z-Transform Method

This method has been discussed in detail in reference 2 where it is shown that one may obtain the actual system response between the sampling instants. Thus, the applications of this method to indicate the hidden oscillations is very straightforward and simple.

For instance, if hidden oscillations are known to exist in the system (from the impulsive response of $G(s)$ and the period T), then one can readily apply the modified z-transform to obtain the actual response between sampling instants in the form of the following integral:[2]

$$C(n,m)T=1/2\pi j\int_{\Gamma}\frac{G^*(z,m)}{1+G^*(z)}\,R^*(z)z^{n-1}dz \tag{4}$$

and the time t is related to (n) and (m) in the following expression:

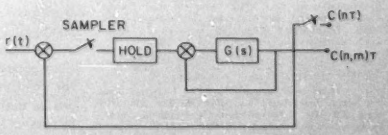


Fig. 6. Sampled-data control system with multiple loop

$$t=(n-1+m)T \tag{4a}$$

where Γ is a path of integration in the z-plane which includes the singularities of the integrand. Since m is considered as constant in the integration, the evaluation of this integral is simple as shown in reference 2. To obtain the total response let m continually vary between 0 and 1; m equals unity yields the response at the sampling instants.

To obtain the maximum of these oscillations the ordinary rules of differentiation can be applied in the following expression:

$$\frac{\partial c(n,m)T}{\partial m}=0 \tag{5}$$

From the discussion on the impulsive response indicated earlier, the severe part of these oscillations usually appear either in the first or second sampling period. Therefore the solution of equation 5 yields the value of m for the maxima in the first or second interval or in any other sampling interval, depending on where the maximum lies.

This method is exemplified by two examples which illustrate the simplicity of application.

In conclusion, it may be noted that the form of $G(s)$ reveals the hidden oscillations as illustrated earlier, and the application of the modified z-transform method gives the magnitude and occurrences of these oscillations. (They are termed "hidden" in this paper because the z-transform method fails to indicate their existence, as explained in detail in Appendixes I and II.)

Initial and Final Value Theorems in the Form of Modified z-Transform

In reference 11, the initial and final value theorems for z-transform has been indicated. These theorems are extended to the modified z-transform, and quickly yield the initial and final value of the continuous response without evaluating the inverse transform, as shown in the following.

The initial value theorem is

$$\lim_{t\to 0} C(n,m)T=\lim_{\substack{z\to\infty\\ m\to 0}} zC^*(z,m) \tag{6}$$

The response for the first sampling period can be obtained from the initial value theorem, as follows:

$$\lim_{\substack{0\leqslant t\leqslant T\\ 0\leqslant m\leqslant 1}} C(n,m)T=\lim_{z\to\infty} zC^*(z,m) \tag{7}$$

The final value theorem is

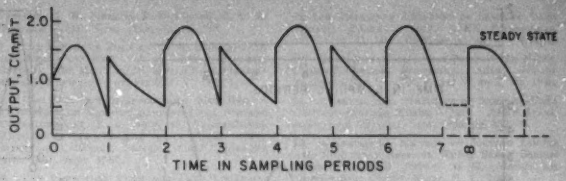


Fig. 7. Response of an open-loop sampled-data system illustrated in Appendix I

$$\lim_{\substack{n\to\infty\\ 0\leqslant m\leqslant 1}} C(n,m)T=\lim_{z\to 1}\frac{z-1}{z}C^*(z,m) \tag{8}$$

if the limit exists

It may be noted that the foregoing relation yields the output steady-state ripple.[4]

Conclusions

It is evident from the material presented that the choice of applying the z-transform method or other method depends mainly on the systems characteristics as well as on the sampling period.

The impulsive response of the open-loop system, compensated or otherwise, will yield sufficient indication as to hidden oscillations. The location of the complex poles of $G(s)$ in the s-plane is a criterion for existence of hidden oscillations. In particular the magnitude of the imaginary part of the complex poles of $G(s)$ and the value of the sampling period indicate the uncertainty of the use of the z-transform method.

It is shown that the modified z-transform method is a powerful tool for detecting these hidden oscillations, which the ordinary z-transform fails to indicate.

It is expected that this paper will eventually aid the designer of sampled-data systems in the choice of applying the z-transform method or any other method to arrive at his objectives.

Appendix I. Illustrative Example

1. Assume an open-loop sampled-data system whose $G(s)$ is as follows:

$$G(s)=\frac{s^3+(\pi+2\alpha)s+\pi(\pi+\gamma)+\alpha^2}{(s+\gamma)[(s+\alpha)^2+\pi^2]} \tag{9}$$

The z-transform[11] and the modified z-transform[3] of equation 9 can be obtained as follows for $T=1$

$$G^*(z)=\frac{z}{z-e^{-\gamma}}+0 \tag{10}$$

$$G^*(z,m)=\frac{e^{-m\gamma}}{z-e^{-\gamma}}+e^{-\alpha m}\times \frac{z\sin m\pi+e^{-\alpha}\sin(1-m)\pi}{(z+e^{-\alpha})^2}$$

$$=\frac{e^{-m\gamma}}{z-e^{-\gamma}}+e^{-\alpha m}\frac{\sin m\pi}{z+e^{-\gamma}} \tag{11}$$

It is to be noticed that the modified z-transform has more poles than $G^*(z)$ and that this is an indication of hidden oscillations.

If the response of the system to step input is computed for values of $\gamma=1$, $e^{-\alpha}=0.98$, the difference between the z-transform information and the modified z-transform is as shown in Fig. 7.

Furthermore, $G(s)$ in this case becomes:

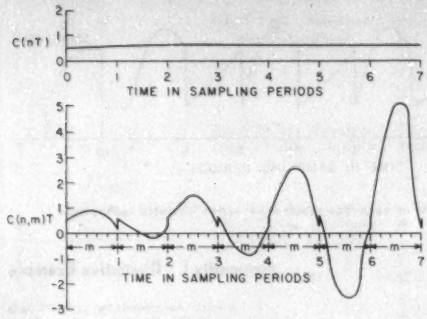
$$G(s)=\frac{1}{s+\gamma}+\frac{\pi}{(s+\alpha)^2+\pi^2} \tag{12}$$

and if $e^{-\alpha T}$ is larger than unity or α is negative, then $G(s)$ itself is unstable which makes the system between sampling instants unstable and the form of $G(s)$ will indicate this.

The initial and final values of the response

Table I. Short Table of z-Transform and Modified z-Transform of Certain Forms of G(s)

G(s)	G*(z)	G*(z,m)
$\frac{k\pi/T}{s^2+(k\pi/T)^2}$	0	$\frac{\sin mk\pi}{z+(-1)^{k+1}}$
$\frac{\pi/T}{s^2+(\pi/T)^2}$	0	$\frac{\sin m\pi}{z+1}$
$\frac{\pi/T}{(s+\alpha)^2+(\pi/T)^2}$	0	$e^{-\alpha mT}\left[\frac{e^{-\alpha mT}\sin m\pi}{z+e^{-\alpha T}}\right]$
$\frac{(s+\alpha)}{(s+\alpha)^2+(\pi/T)^2}$	$z/z+e^{-\alpha T}$	$e^{-\alpha mT}\frac{\cos m\pi}{z+e^{-\alpha T}}$
$\frac{s+a}{(s+\alpha)^2+(\pi/T)^2}$	$\frac{z}{z+e^{-\alpha T}}$	$e^{-\alpha mT}\left[\frac{T(a-\alpha)}{\pi}\sin m\pi+\cos m\pi\right]$



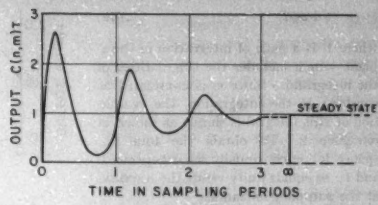


Fig. 8 (left). Response of a system illustrated in Appendix I, Item 2

Fig. 9 (above). Response of the systems illustrated in Appendix II

can be obtained by applying relations 6 and 8 as follows:

$$\lim_{\substack{t\to 0\\ m\to 0}} C(n,m)T = \lim_{\substack{z\to\infty}} zG^*(z,m)z/z-1 = 1.0 \tag{13}$$

$$\lim C(n,m)T = \lim\,(z-1/z)G^*(z,m)(z/z-1) = n\to\infty,\ 0\leqslant m\leqslant 1\ \ z\to 1 = \varepsilon^{-m\gamma}/1-\varepsilon^{-\gamma}+\varepsilon^{-\alpha m}\sin m\pi/1+\varepsilon^{-\alpha} \tag{14}$$

2. Assume the same system discussed is to operate as sampled-data feedback system as shown in Fig. 1. Thus the z-transform of the output for step input is given

$$C^*(z) = \frac{z}{z-1}\times\frac{G^*(z)}{1+G^*(z)} \tag{15}$$

where

$$G^*(z) = \frac{z}{z-\varepsilon^{-\gamma}} \tag{16}$$

and $C^*(z,m)$ is given

$$C^*(z,m) - \frac{G^*(z,m)}{1+G^*(z)}\times\frac{z}{z-1} \tag{17}$$

If the response is calculated for $\gamma=1$, $\varepsilon^{-\alpha}=1.5$, then it is to be noticed that this response at the sampling instants is stable, while that in between sampling instants is unstable. This is indicated in Fig. 8. The cause of this instability is that the impulsive response of $G(s)$ in equation 12 is unstable for the choice of such a value of α.

Appendix II. Illustrative Example with Feedback

Assume a system having a transfer function $G(s)$ of the following form:

$$G(s) = \frac{27.8(s+1)}{s[(s+3)^2+(2\pi)^2]},\ T=1 \tag{18}$$

The corresponding $G^*(z)$, computed using Table I of reference 11, is

$$G^*(z) = \frac{0.543\times z}{(z-0.05)(z-1)} \tag{19}$$

The modified z-transform $G^*(z,m)$, found from Table I, is as follows:

$$G^*(z,m) = 0.572\left[\frac{1}{z-1} - \varepsilon^{-2m}\times\frac{(\cos 2\pi m - 7.27\sin 2\pi m)}{z-\varepsilon^{-3}}\right] \tag{20}$$

The output $C(n,m)T$ for a step input, found by use of the inverse z-transform, is

$$C(n,m)T = \frac{1}{2\pi j}\int_\Gamma \frac{G^*(z,m)}{1+G^*(z)}\times\frac{z}{z-1}z^{n-1}dz \tag{21}$$

Inserting the foregoing values of $G^*(z,m)$ and $G^*(z)$ in equation 21 yields

$$C(n,m)T = \frac{0.572}{2\pi j}\times\int_\Gamma \frac{z-0.05-\varepsilon^{-3m}(z-1)(\cos\pi m - 7.37\sin 2\pi m)}{\Gamma[(z-0.05)(z-1)+0.543z][z-1]}\times z^n dz \tag{22}$$

To find the maximum output form:

$$\frac{\partial C(n,m)T}{\partial m} = 0 \tag{23}$$

This gives

$$\frac{\partial C(n,m)}{\partial m} = 0 = 0.572\varepsilon^{-3m}(-15.53\sin 2\pi m + 48.6\cos 2\pi m)\times\frac{1}{2\pi j}\int_\Gamma \frac{z^n dz}{(z^3-0.507z+0.05)} \tag{24}$$

Since the integral is never zero, except for $n=0$ which is a trivial case,

$$48.68\cos 2\pi m - 15.53\sin 2\pi m = 0 \tag{25}$$

or

$$\sqrt{48.6^2+15.53^2}\cos(2\pi m+\theta) = 0,\qquad \tan\theta = \frac{15.53}{48.6},\ \theta = 17.8^\circ \tag{26}$$

This gives

$$2\pi m + 17.8^\circ = \pi/2,\ m_{max} = 0.2005 \tag{27}$$

Substituting this value of m_{max} in $C(n,m)T$ gives

$$C_{max}(n,m)T\,\frac{1}{2\pi j}\int_\Gamma \times\frac{(4.63z-3.68)z^n dz}{(z-1)(z-0.135)(z-0.372)} \tag{28}$$

This yield for $n=1$

$$C_{max} = 2.64$$

and for $n=2$

$$C_{max} = 1.885 \tag{29}$$

and for $n=3$

$$C_{max} = 1.37 \tag{30}$$

The plot of $C(n,m)T$ is shown in Fig. 9. It is noticed that the maxima coincide with the calculated values.

Comparing the actual response using the modified z-transform with the response at the sampling instants obtained by using the z-transform, shows the hidden oscillations.

The fact that the impulsive response is of oscillatory form and T is larger than half the period of oscillation is an indication that the z-transform cannot be simply applied.

The initial and final values of the response can be obtained by applying relations 6 and 8 to $C^*(z,m)$ in equation 22.

For the initial value

$$\lim_{\substack{t\to 0\\ m\to 0}} C(n,m)T = \lim_{\substack{z\to\infty}} zC^*(z,m) = 0 \tag{31}$$

For the final value

$$\lim\ C(n,m)T = \lim\frac{z-1}{z}C^*(z,m) = 1 \tag{32}$$

$n\to\infty,\ 0\leqslant m\leqslant 1\ \ z\to 1$

Thus, in this case the final value is unity for all values of m.

References

1. The Pulse Transfer Function and Its Application to Sampling Servosystems, R. H. Barker, *Proceedings*, Institute of Electrical Engineers, London, England, vol. 99, 1952, p. 1.

2. Synthesis and Critical Study of Sampled-Data Control Systems, E. I. Jury, Electronics Research Laboratory Report, University of California, Berkeley, Calif., series 60, issue 136, 1955.

3. The Design of Sampled-Data Feedback

Systems, Gladwyn V. Lago, John G. Truxall. *AIEE Transactions*, vol. 73, pt. II, Nov. 1954, pp. 247–53.

4. Root-Locus Method of Pulse Transfer Function for Sampled-Data Control Systems, M. Mori. *Automatic Control*, Japan, no. 1, 1955, pp. 5–12.

5. Analysis of Errors in Sampled-Data Feedback Systems, Jack Sklansky, J. R. Ragazzini. *AIEE Transactions*, vol. 74, pt. II, May 1955, pp. 65–71.

6. Extension of Continuous-Data System Design Techniques to Sampled-Data Control Systems, G. W. Johnson, D. P. Lindorf, C. G. A. Nordling. *Ibid.*, Sept. 1955, pp. 252–63.

7. The Effect of Pole and Zero Locations on the Transient Response of Linear Dynamic Systems, John Milligan, Jr. *Proceedings*, Institute of Radio Engineers, New York, N. Y., May 1949.

8. Table of Z-Transform and Modified Z-Transform of Various Sampled-Data Systems Configuration, E. I. Jury, G. Farmanfarma. Electronics Research Laboratory Report, University of California, Berkeley, Calif., series 60, issue 136a, Sept. 1955.

9. The Effect of Pole and Zero Locations on the Transient Response of Sampled-Data Systems, Eliahu I. Jury. *AIEE Transactions*, vol. 74, pt. II, Mar. 1956, pp. 41–48.

10. Communications in the Presence of Noise, Claude E. Shannon. *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 37, no. 1, Jan. 1949.

11. Analysis and Synthesis of Sampled-Data Control Systems, Eliahu I. Jury. *AIEE Transactions*, vol. 73, pt. I, Sept. 1954, pp. 332–46.

12. A Simplified Method of Solving Linear and Nonlinear Systems, Rubin Boxer, Samuel Thaler. *Technical Report 55–44*, Rome Air Development Center, Rome, N. Y., 1955.

Modernization of the Long Island Rail Road Passenger-Car Fleet

P. H. HATCH
FELLOW AIEE

THE rehabilitation and modernization of any organization—industrial, transportation, commercial—is usually of interest. When such an organization is a railroad handling 285,000 passengers per day, its rehabilitation and modernization become a matter of more than usual interest. Such is the case of the Long Island Rail Road.

It emerged from receivership on August 12, 1954, as a redevelopment corporation, committed to do exactly what the term implied: bring its property and operations up to modern standards.

Its equipment ownership included both locomotives and cars. The latter far outnumbered the locomotives and were of two general types: cars used in locomotive-hauled trains and electric cars made up into multiple-unit trains. As of August 12, 1954, Long Island ownership consisted of 18 steam locomotives, 65 diesel-electric locomotives, 10 electric locomotives, 322 locomotive-hauled cars, and 999 multiple-unit electric cars. Of this total of 1,321 cars, only a very small proportion could make any claim to being modern. Hence the rehabilitation of this fleet became first on the list of projects to be undertaken.

The rehabilitation consisted of two equally important parts: (1) the ordering of 220 new cars, not including two new diesel rail cars; and (2) the modernization of 696 of the existing cars which were sound structurally but which had not kept pace with developments in seating, lighting, ventilation, and other interior arrangements, and which also required upgrading of electric equipment. While buying new cars and modernizing old cars would seem to have little in common, actually the two were intentionally related to a considerable degree in the interests of standardization and economy.

In both new and older cars, no component plays a more important part or is used for a wider variety of functions than the electric apparatus. This paper will outline the over-all car rehabilitation and modernization plan, with particular reference to electrical features.

New Cars

In New York City and other large metropolitan areas, the problem in handling heavy suburban traffic is getting more passengers into and out of the city. Usually terminal track space has been limited, which restricts the number of trains which can be run. The only practicable way to accommodate more passengers, therefore, is to handle more of them per train and per car. This had led to the development of two special types of suburban car: the multiple-level car such as the so-called gallery car and double-deck car; and the long, single-deck car with 3-passenger and 2-passenger seating.

After careful consideration, the Long Island decided to purchase the latter type of car which had worked well on other railroads entering New York and which, in small quantity, it had itself operated. It was decided further to use the same basic car in both locomotive-hauled trains and multiple-unit electric trains, and to use and combine items of car equipment already developed and on the market. It was also decided to omit anything which did not directly contribute to safety, comfort, and economy. The result was a completely modern, high-capacity car, comfortable to ride in and pleasing to the eye, which could be purchased at a cost low enough to permit buying 220 cars with the money originally allotted for 184. Accordingly, two orders were placed for a total of 220 cars: one for 80 cars for locomotive train service and one for 140 cars for multiple-unit train service.

A factor which contributed considerably to the cost situation arose because Long Island multiple-unit trains are made up of head-end motor cars (motor cars having operating compartments at each end and fitted with automatic train speed control), so-called motorized trailers (motor cars without operating compartments designed solely for use in the middle of trains), and straight trailer cars (non-motor cars without operating compartments). The value of such cars is in the same order, i.e., the head-end cars are the most expensive, the motorized trailers are next, and the straight trailers are the least expensive. Hence in placing the orders for 140 multiple-unit cars, only 22 head-end motor cars were specified with the balance consisting of 74 motorized trailers and 44 straight trailers. This division of the orders kept the first cost down for 140 of the new cars, and also made full use of existing head-end motor cars.

General Details of the New Cars

The basic Long Island passenger car, both for locomotive-hauled and multiple-unit trains, has a length of 82 feet, 5 inches, between pulling faces of the couplers, an over-all width of 10 feet, $5^1/_8$ inches, and an over-all height of 13 feet, $7^1/_8$ inches.

Paper 56-979, recommended by the AIEE Land Transportation Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Fall General Meeting, Chicago, Ill., October 1–5, 1956. Manuscript submitted May 31, 1956; made available for printing July 20, 1956.

P. H. Hatch is with the Long Island Rail Road Company, Richmond Hill, N. Y.

Fig. 1. Locomotive-hauled car

Fig. 2. Multiple-unit head-end car

Fig. 3. End view of multiple-unit head-end car

Fig. 4. View of typical car interior

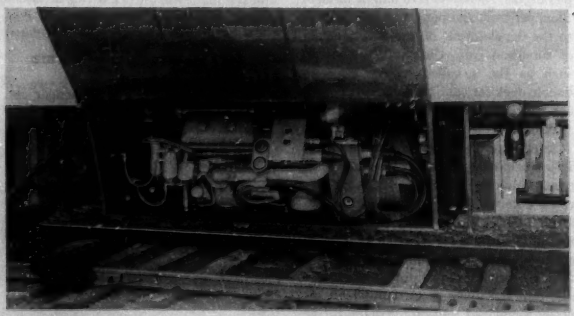


Fig. 5. Diesel-engine–generator set for power supply on locomotive-hauled car

Light weights, in working order, are 107,300 pounds for the locomotive-hauled car, 132,100 pounds for the multiple-unit head-end motor car, 129,500 pounds for the multiple-unit motorized trailer car, and 107,000 pounds for the multiple-unit trailer car.

Seating capacity of the locomotive-hauled car and the head-end multiple-unit motor car, both of which types have toilets, is 120 passengers; seating capacity of the other two types of car is 123 passengers. Thus, the 220 new cars provide total seating capacity for 26,754 persons.

All cars are constructed of high-tensile low-alloy steel, and have vestibules at each end, so-called "turtle-back" roofs, and rubber draft gear. They are mounted on modified all-coil spring, 4-wheel cast-steel trucks. (The modification consists of single elliptic bolster springs to eliminate the use of hydraulic shock absorbers.) All cars have center-trough fluorescent lighting, air conditioning of 8 tons' capacity, specially designed and constructed 3- and 2-passenger seats, anodized aluminum baggage racks, steam or electric side-wall heating, and rubber tile floors. Interior decoration is stainless steel with a red stripe over the windows and painted head-lining. Exterior decoration is in accordance with the railroad's standard painting scheme of green roofs, dark gray sides, and black trucks and underframe equipment. Letters and numbers are red or light gray in color.

Air-brake equipment on all cars is of the latest type. On locomotive-hauled cars it is Westinghouse air brake schedule *D-22-P*, on head-end multiple-unit elec-

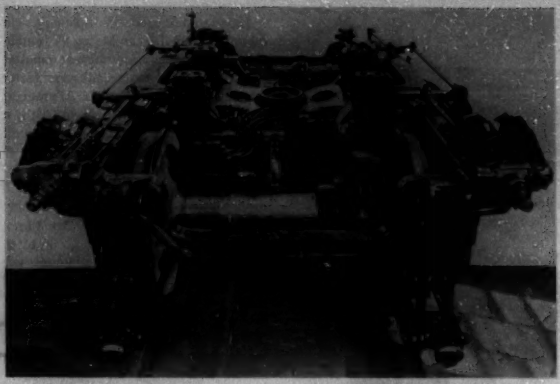


Fig. 6. Motor truck with traction motors, flexible couplings, and gear units

tric cars it is schedule *D-22-MDE*, and on motorized and straight trailer cars it is schedule *D-22-TDE*. All of these schedules use the *D-22* control valve which produces highly uniform braking effort under varying conditions. All cars have two truck-mounted brake cyclinders on each truck with automatic and manual slack adjusters. The multiple-unit cars have electropneumatic brake control and the head-end multiple-unit cars have automatic 3-speed coded train control superimposed on the air-brake system.

Figs. 1 and 2 show, respectively, a locomotive-hauled car and a multiple-unit head-end car. The similarity of the two types is readily seen. Fig. 3 shows an end view of one of the multiple-unit head-end cars. Fig. 4 shows an interior view which is typical of all 220 cars. The circular badge plate with number in the center denotes the number of new cars delivered up to and including the car on which the badge is applied.

Electrical Details of the New Cars

Locomotive-Hauled Cars

Since these cars were to be used principally in commuter service where the runs are comparatively short and the stops frequent, adequate power supply was a serious problem, especially since air conditioning was involved. Existing cars having only electric lighting to contend with required constant attention to battery charging at either one or both ends of the runs since the axle generators had difficulty in keeping the batteries sufficiently charged. The stepping up of the power demand on the new cars virtually ruled out the axle generator-storage battery system as a practicable method. It was decided to install so-called under-car power plants which would make the cars independent of speed, number of stops, and distance run, and would eliminate the need for installation of expensive terminal stand-by facilities.

The use of diesel-engine-driven alternators on the 80 locomotive-hauled cars makes this the largest installation of its kind in the east. These sets, which supply power for lighting, electromechanical air conditioning, and the electric water cooler, have a 35-horsepower diesel engine direct-connected to a 25-kw a-c generator. Power supplied is 220 volt, 60 cycle, 3 phase, a-c, with provision for supplying 32-volt d-c power for miscellaneous lights, battery charging, and control. Three degrees of car heating and three degrees of car-cooling control are provided. The water cooler has a standard 110-volt single-phase motor, and to supply this a single-phase 220/110-volt transformer is connected across one phase of the alternator supply.

In the event of failure of the engine-generator set, the car may be jumpered to an adjoining car in which case a-c power for lighting and ventilation but not cooling will be available.

Fig. 5 illustrates the diesel-engine–generator set mounted beneath the car.

Multiple-Unit Cars

Basic Factors

The electrified territory in which the multiple-unit cars operate utilizes a 650-volt d-c overrunning third rail.

In considering the new multiple-unit cars, it is necessary to discuss briefly from an engineering viewpoint the background of multiple-unit operation on the Long Island into which the new cars had to fit.

As previously stated, Long Island ownership of multiple-unit cars as of August 12, 1954, was 999. Of this total, 751 were motor cars and 248 were trailer cars. They were made up in trains with a motor-car-to-trailer-car ratio of not less than 70 to 30. Furthermore, only seven of the 751 motor cars had four traction motors; all the rest of the motor cars had two motor equipments, one truck being a motor truck and the other being a trailer truck. Total car weights were divided approximately 57 per cent on motor trucks and 43 per cent on trailer trucks. It can be seen that a great deal of adhesion weight was thus unutilized.

From the operating standpoint, the ideal multiple-unit train consists of all motor cars with all axles motorized. Including trailers with motor cars introduces complications in operation and fails to take advantage of all available adhesion weight. When this is done with motor cars having two motor equipments, still less advantage of available adhesion

Fig. 7. Westinghouse type 1453-A traction motor

Fig. 8. General Electric type 1240-C1 traction motor

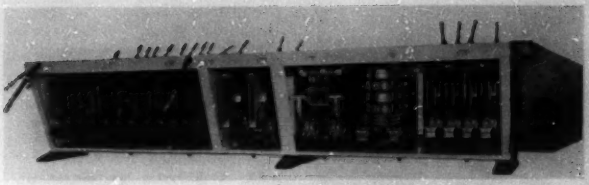


Fig. 9. Arc box side of Westinghouse control box with covers removed

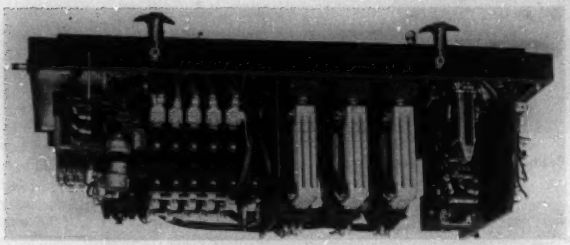


Fig. 10. Front view of General Electric control group with cover removed

weight is realized. This has the effect of putting a premium on good rail conditions and tends to limit acceleration rates.

In planning the new multiple-unit cars, economic considerations dictated the continuance of the motor-car–trailer-car operation. It was, however, decided to go to four motor equipments for all motor cars and the operating results to date have more than justified such a move.

With the older equipment, the number of motor cars in a train must be at least 70 per cent of the total cars; with the new equipment this ratio can be reduced to 60 per cent. Advantage has not been taken of this so far since it was desired to operate the new cars in various combinations with the old cars and 70 per cent is still the governing figure under these conditions.

Current limit relays are set to give an acceleration rate of 1.0 mph (miles per hour) per second for Long Island multiple-unit trains with a 70 to 30 motor–trailer ratio. An attempt had previously been made to increase this to 1.25 mph per second on some of the trains with the then more modern motor cars. Considerable difficulty was experienced in keeping motor cars which were so adjusted together in the same train. As the result, cars with both acceleration rates were found in the same train, which was far from desirable. To rectify this, the acceleration rate for all motor cars, including the new ones, is now adjusted on the basis of 1.0 mph per second for trains with 70 to 30 ratio of motor and trailer cars. It is obvious that for a greater motor–trailer ratio in a given train, the resulting acceleration rate will be higher.

With a view toward the future, the new motor cars have capacity and adjustment so the acceleration rate may be increased to 1.25 mph per second for trains made up of new equipment with a 60 to 40 motor–trailer ratio. It is understood, of course, that actual operation at these rates will be dependent upon substation and distribution system capacity.

Having discussed multiple-unit car adhesion, 2-motor and 4-motor equipments, and acceleration rates, as a matter of interest the average adhesion factor (required tractive effort divided by weight on driving wheels) was calculated for a typical 10-car Long Island multiple-unit train of the older cars with 70 to 30 motor–trailer ratio, two motor equipments, and an acceleration rate of 1.0 mph per second on straight, level track. The same factor was then calculated for a 10-car train of new cars with a 60 to 40 motor–trailer ratio and 1.25-mph per second acceleration rate on straight, level track. These respective adhesion factors are 11.9 and 10.1 per cent. The lower adhesion factor for the new cars under more severe operating conditions illustrates the adhesion margin built into them.

Dynamic braking was considered for the motor cars, but the operation of trailer cars in trains precluded its use because of traction motor heating. A further consideration was the required operation of the new motor cars with the older motor cars.

Equipment Details

Electric equipment for both propulsion and auxiliary purposes was available from both major manufacturers. Furthermore, the same or basically similar apparatus had extensive service experience behind it on other railroads as well as on the New York City Transit Authority properties.

It was decided to purchase half of the electric equipment from one manufacturer and half from the other with the provision that traction motors, flexible couplings, and gear assemblies be interchangeable as units, and with the further provision that all new motor cars would operate in multi-

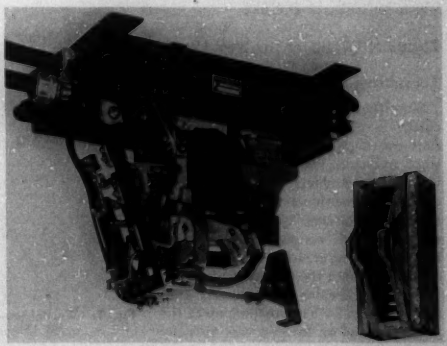


Fig. 11. One type of line switch with cover and auxiliary arc chute removed

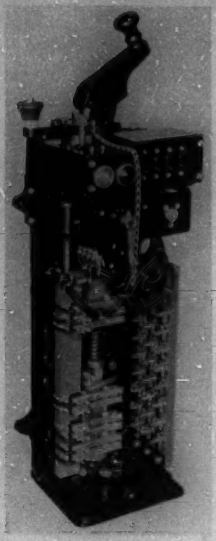
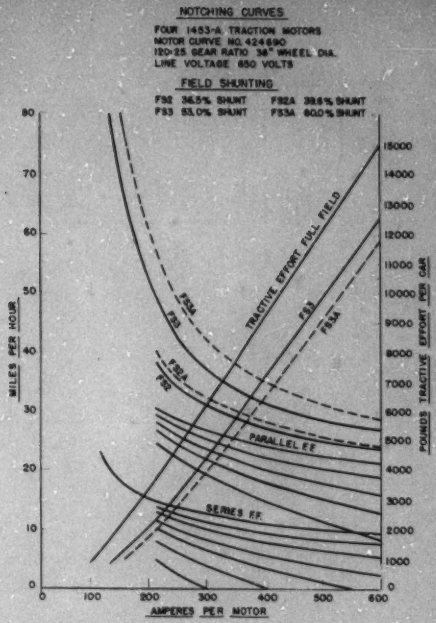


Fig. 12. Master controller with cover removed

Fig. 14. Notching curves for four Westinghouse type 1453-A traction motors



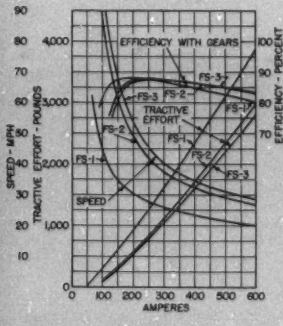


Fig. 13. Characteristic curves for General Electric type 1240-C1 traction motor

ple with each other and with all types of existing motor cars.

Each motor car, either head-end or motorized trailer, has four 100-horsepower truck-mounted high-speed traction motors. Associated with these on each car are a high-capacity line switch, switch group arranged for automatic acceleration, reverser, ribbon-type resistor units, air compressor, 650-volt motor-generator set for battery charging and 32-volt power supply, 32-volt motor alternator for 120-volt a-c fluorescent-light supply, 300-ampere-hour storage battery, necessary knife switches, fuse boxes, etc.

The straight trailer cars have the same battery-charging and fluorescent-light-supply motor-generator sets, storage battery, and other auxiliary equipment as the motor cars. Both motor and trailer cars have the same air-conditioning equipment with three degrees of heating and three degrees of cooling control. A feature of both motor and trailer cars is an exceptionally short, equalizer-supported third-rail shoe beam that is standard in all positions on all cars.

It will be noted that third-rail shoes are applied to the new trailer cars. This was done because the auxiliary power demands of these cars would overload the existing bus line with which all multiple-unit cars are equipped. Nominal rating of this bus line is 500 amperes and it handles propulsion power over third-rail gaps as well as auxiliary power.

Fig. 6 shows one of the motor-car trucks with frame-mounted traction motors. Note the flexible coupling and gear unit. Motors are geared for a maximum speed of 101 mph with 36-inch wheels. With a maximum allowable operating speed of 65 mph, there is ample margin in traction-motor armature speed. This, together with the adhesion margin previously mentioned, makes the use of slip relays unnecessary.

Figs. 7 and 8 show, respectively, the Westinghouse *1453-A* and General Electric *1240-C1* traction motors.

Figs. 9 and 10 show two types of control groups. The front of the main control assembly, Fig. 9, includes, left to right, resistor and motor combination switches, line switch, relay panel, drum-type reverser, and motor field shunt switches. In the housing at the right end are the inductive field shunts and resistors. In Fig. 10 are one side of the electropneumatic cam-operated controller, the three motor combination switches, and a relay panel. On the end are mounted miscellaneous small resistor tubes. Not shown are the field shunt switches, drum-type reverser, and field shunt relay contained in a second box.

Fig. 11 shows the line switch used with the control equipment of Fig. 10. Fig. 12 shows the master controller. Fig. 13

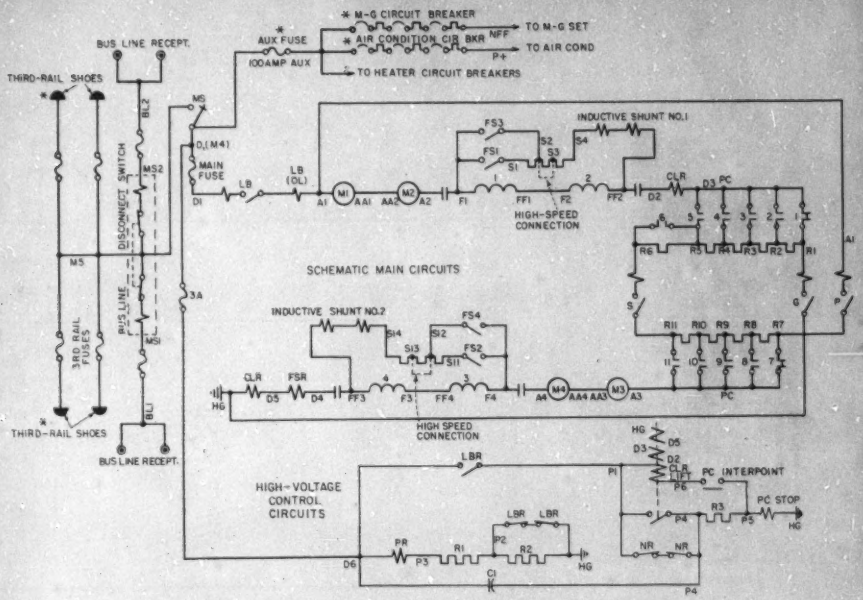


Fig. 15. Main-circuit schematic diagram for General Electric motor car

gives characteristic curves for the General Electric traction motor and Fig. 14 shows notching curves for a Westinghouse-equipped car. Figs. 15 and 16 show, respectively, schematic diagrams of main circuits on the two types of motor cars.

Modernization of Existing Cars

General Details

When the Long Island emerged as a redevelopment corporation, its passenger-car fleet consisted of 322 locomotive-hauled cars and 999 multiple-unit electric cars. Since one of the objectives of this corporation was the modernization of its passenger-car fleet, plans were made to rehabilitate and modernize 217 of the locomotive-hauled cars and 479 of the multiple-unit electric cars in the company shops.

Cars to be modernized were stripped down to the bare structures. Necessary repairs and reinforcements were made to car bodies and trucks were completely dismantled and overhauled. Interiors were modified and fitted so that the general appearance is the same as that of the new passenger cars.

Modernized cars have two rows of incandescent light fixtures which produce 20–25-foot-candle intensity at the reading plane, five overhead paddle-type circulating fans, Long Island standard 3-passenger and 2-passenger seats, anodized aluminum baggage racks, side-wall steam or electric heating, and rubber tile floors. The interior and exterior painting scheme is the same as for the new cars. Seats, baggage racks, heater boxes, and flooring, and many other details, are identical with similar equipment in the new cars.

By the use of 3-2 passenger seating, the capacity of a typical modernized car has been increased from 78 to 89 passengers, or from 80 to 91 passengers, depending on the type of car. Typical light weight, in working order, for the motor cars is 115,500 pounds and for the trailer cars is 74,000 pounds.

In addition to overhaul of trucks on all cars, 43 cars will have roller journal bearings installed.

Electrical Details

In a majority of the existing cars, the electric propulsion equipment was somewhat antiquated, with a high maintenance and mediocre performance record. Even before details of the car modernization plan had finally been settled, Long Island specifications for armature and field coils were drawn up and a program for rewinding traction motors was started. Insulation of armatures and fields is between grades *B* and *H*, and when sufficient motors are completed, a material improvement in performance is expected with the upgraded insulation.

Other traction motor improvements include all-weather covers on certain motors, installation of roller bearings and larger axle bearings on other motors, and revision of commutating-pole fields on others to improve electrical stability. In addition to these changes, is a reapportionment of resistance steps and recalibration of accelerating and limit relays to smooth out train handling and to reduce undesirable effects of excessive current conditions.

A weak point of many of the older cars was the line switch (circuit breaker) which was subject to occasional blowups, frequently accompanied by 650-volt third-rail power entering the control circuits. Also, these switches did not provide suffi-

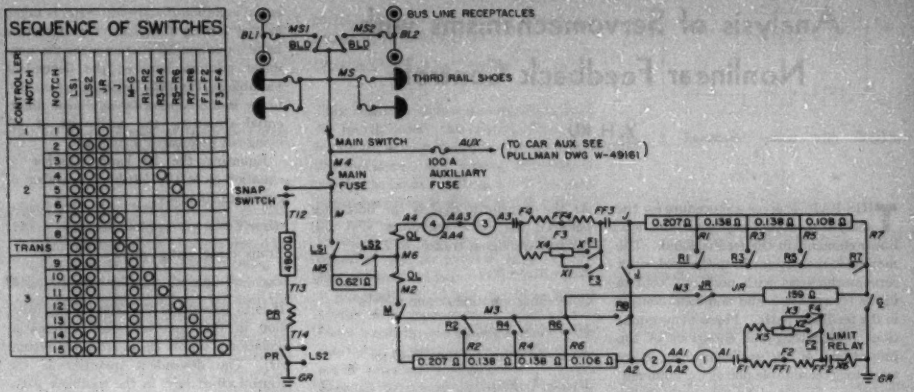


Fig. 16. Main-circuit schematic diagram for Westinghouse motor car

cient protection against damage when failures of equipment or wiring occurred. It was thus considered of primary importance to provide completely modern and efficient line switches as part of the electrical modernization.

New line switches are being installed on a number of the cars. These are either single-contact switches, or double-contact switches to fit in with the control arrangement on certain cars. It is planned to modernize existing line switches on certain other cars to make them as nearly the equivalent to the new switches as possible.

As the interrupting ability of the new type of line switches on short circuit is approximately 30,000 amperes at 750 volts d-c, they are capable of successfully opening a short circuit directly adjacent to a high-capacity substation. This has been confirmed by actual test on the railroad.

Another item requiring improvement was the 16-volt control train line in a majority of the older cars. All cars have now been changed over to utilize 32-volt control power.

A further item of improvement has been the replacement of control batteries with batteries of approximately double the ampere-hour capacity. This was undertaken on the head-end cars first and is being extended to other cars on a program basis.

It can readily be understood that electrical rehabilitation on as broad a scale as this, and applied to hundreds of cars, is a large undertaking. Every effort is being made to release modernized cars incorporating all of these improvements, but due to material delivery and other considerations this has not always been possible in its entirety. Hence a great many of the improvements are being applied to cars at every available opportunity and not just to cars being modernized. The result will be the same in the end and better economy and performance can begin to be realized in the meantime.

Maintenance

By no means least in a program of purchase of new cars and rehabilitation and modernization of old cars in a large fleet, are provisions for adequate supporting maintenance and so $1,500,000 has been allotted for this factor.

The principal projects to be included which are related to cars are a modern wheel shop with up-to-date machinery, rearrangement of shop areas and installation of new equipment for traction motor and generator repairs, a wheel grinder to recondition wheels without removal from the cars, new testing apparatus, and a new multiple-unit car-inspection shop. This latter will be the most modern installation of its kind in the United States.

Conclusions

The modernization of the passenger-car fleet is not the entire story of the modernization of the Long Island Rail Road. Dollar-wise, however, and from the standpoint of impact on the public, it is the most important single segment of the entire program. It is especially gratifying to note the favorable response to the new cars and to the rebuilt older cars and everything possible is being done to expedite the work so that the railroad's passenger-car fleet as a whole will be brought up to the requirements of the modernized Long Island.

Analysis of Servomechanisms with Nonlinear Feedback Control

Y. H. KU
FELLOW AIEE

THIS PAPER gives a procedure for the analysis of servomechanisms with nonlinear elements in the feedback link. The method is then extended to the analysis of servomechanisms with nonlinear control in the forward link and nonlinear control in the feedback path. Phase-plane equations representing the equations of the projections of an n-dimensional space trajectory on suitable planes are obtained. These can be solved by the methods detailed in previous papers.[1-3]

Let the feedback circuit be as shown in Fig. 1. In the forward branch is a linear network with its transfer function $G_1(s)$, where $G_1(s)$ is given in terms of Laplace transforms with s denoting the complex variable. In the feedback path are two networks, one linear and the other nonlinear. The linear network has a transfer function $G_2(s)$. The nonlinear control network is denoted by G_{N2} and its characteristics will be defined later.

Nomenclature

The notation of Fig. 1 and the corresponding transforms are defined in the following.

$r(t)$ = reference input to the system
$R(s)$ = Laplace transform of $r(t)$
$b(t)$ = feedback signal
$B(s)$ = Laplace transform of $b(t)$
$e(t) = r(t) - b(t)$ = actuating signal
$E(s)$ = Laplace transform of $e(t)$
$c(t)$ = output of the system
$C(s)$ = Laplace transform of $c(t)$
$m(t)$ = control effort in the feedback link
$M(s)$ = Laplace transform of $m(t)$
$m_1(t)$ = control effort in the forward link
$M_1(s)$ = Laplace transform of $m_1(t)$

Laplace Transform Equations

Since the networks G_1 and G_2 are linear, the following equations can be written:

$$G_1(s)E(s) = C(s) \tag{1}$$

$$G_2(s)C(s) = M(s) \tag{2}$$

Paper 56-954, recommended by the AIEE Feedback Control Systems Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Fall General Meeting, Chicago, Ill., October 1-5, 1956. Manuscript submitted October 20, 1955; made available for printing August 1, 1956.

Y. H. Ku is with the Moore School of Electrical Engineering, University of Pennsylvania, Philadelphia, Pa.

As the transform $E(s)$ is the difference between the input transform $R(s)$ and the feeback signal transform $B(s)$,

$$R(s) - B(s) = E(s) \tag{3}$$

Combining equations 1 and 2 gives

$$M(s) = G_1(s)G_2(s)E(s) \tag{4}$$

Equation 4 can be simplified to

$$M(s) = G(s)E(s) \tag{5}$$

where

$$G(s) = G_1(s)G_2(s)$$

Substituting equation 3 into equation 5 and transposing terms yields

$$G(s)R(s) - M(s) = G(s)B(s) \tag{6}$$

Dividing both sides of the equation 6 by $G(s)$

$$R(s) - M(s)/G(s) = B(s) \tag{7}$$

Equation 7 gives the relation between the control effort $M(s)$ and the feedback signal $B(s)$ for any given input $R(s)$.

Differential Equations in the Time Domain

In writing the foregoing transform equations, the system is assumed as initially quiescent. Let $G_1(s) = 1/(Ts+1)$ and $G_2(s) = 1/s$, as shown in Fig. 2. Then

$$G(s) = G_1(s)G_2(s) = \frac{1}{s(Ts+1)} \tag{8}$$

Substituting equation 8 into equation 7 and inverting term by term gives the following differential equation:

$$r(t) - (Tm'' + m') = b(t) \tag{9}$$

where the double prime and the prime denote the second and first derivatives with respect to time.

The nonlinear network G_{N2} in the feedback path is assumed such as to give the following functional relation between the feedback signal $b(t)$ and the control effort $m(t)$ and its derivatives.

$$g_{n2}(m'', m', m) = b(t) \tag{10}$$

Equation 9, derived from the linear Laplace transform equations, may be considered a second-order linear differential equation if $r(t) - b(t) = e(t)$ is taken as the forcing function. On the other hand, equation 10 characterizes the nonlinear relation between the dependent variable $m(t)$ and the forcing function $b(t)$, hence is a second-order nonlinear differential equation, the coefficient of m'' being not necessarily a constant.

Equating the left-hand member of equation 9 to the left-hand member of equation 10 gives

$$r(t) - (Tm'' + m') = g_{n2}(m'', m', m) \tag{11}$$

Transposing terms gives

$$Tm'' + m' + g_{n2}(m'', m', m) = r(t) \tag{12}$$

It is to be noticed that the forcing function in the final nonlinear differential equation is simply the input to the system $r(t)$. The dependent quantity is the control effort $m(t)$ in the feedback path. The nonlinearities are preserved in the general nonlinear function g_{n2} characterizing the nonlinear control network G_{N2}.

Equation for a Third-Order Mechanism

Consider the third-order servomechanism shown in Fig. 3. Let the transfer function of the forward link be

$$G_1(s) = \frac{1}{s^2(Ts+1)} \tag{13}$$

Assume that in the feedback path there is only a nonlinear controller having the following characteristics:

$$g_{n2}(c'', c', c) = b(t) \tag{14}$$

It is to be noticed that in this simplified network, the output c is equal to the control effort in the feedback path m.

From equation 1,

$$G_1(s)E(s) = \frac{E(s)}{s^2(Ts+1)} = C(s) \tag{15}$$

From equation 3,

$$B(s) = R(s) - E(s) = R(s) - s^2(Ts+1)C(s) \tag{16}$$

Inversion yields

$$b(t) = r(t) - Tc''' - c'' \tag{17}$$

where the triple prime denotes the third derivative with respect to time. Equating the right-hand member of equation 17 to the left-hand member of equation 14 and transposing terms gives the final nonlinear differential equations

$$Tc''' + c'' + g_{n2}(c'', c', c) = r(t) \tag{18}$$

This nonlinear differential equation is of the third order; the output $c(t)$ is the dependent variable. The correction

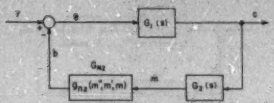


Fig. 1. Block diagram of feedback system with nonlinear control

signal from the feedback path is $b(t)$, which can be obtained by the relation given in equation 14 and is represented by the last term of the left-hand member of equation 18. At any instant, the signal $e(t)$ is given by the difference between the input $r(t)$ and the feedback signal $b(t)$. By solving equation 18, all pertinent information is available. Since r is usually known, dr/dt can be found. By noting the incremental change of $b(t)$ with respect to the chosen time interval, the approximate value of db/dt is also known. Then de/dt, the rate of actuating signal, is given by $dr/dt - db/dt$.

Equation for a Fourth-Order Nonlinear System

If the forward link of the servomechanism has a third-order transfer function given by equation 13, and there is a linear network in the feedback path characterized by the transfer function

$$G_N(s) = \frac{K}{s} \tag{19}$$

the system is of the fourth order. Nonlinear control inserted either in the forward link or the feedback link will make the complete system a fourth-order nonlinear network (Fig. 4).

Combining $G_1(s)$ as given in equation 13 and $G_2(s)$ as given in equation 19 yields

$$G(s) = \frac{K}{s^3(Ts+1)} \tag{20}$$

Substituting equation 20 into equation 5 gives

$$\frac{E(s)K}{s^3(Ts+1)} = M(s) \tag{21}$$

thus

$$E(s) = (1/K)(Ts^4+s^3)M(s) \tag{22}$$

From equation 3,

$$B(s) = R(s) - E(s) = R(s) - (1/K)(Ts^4+s^3)M(s) \tag{23}$$

Inversion yields

$$b(t) = r(t) - (T/K)m'''' - (1/K)m''' \tag{24}$$

where the quadruple prime denotes the fourth derivative with respect to time. Assuming that equation 10 gives the correct control function and combining equations 10 and 24 gives

$$(T/K)m'''' + (1/K)m''' + g_{N3}(m'', m', m) = r(t) \tag{25}$$

This nonlinear differential equation is of the fourth order; the control effort $m(t)$ is the dependent variable. In equation 25, see above, the forcing function is simply the input r, as before. If the nonlinear controller is placed in the forward link instead of in the feedback link, the forcing function in the final differential equation will, in general, involve high derivatives of the input r.

Equations 12, 18 and 25 can be solved by the phase-space method given in references 1 and 2. An example of the solution of a second-order nonlinear differential equation with a nonlinear coefficient on the second derivative is given in reference 3. Examples involving third- and higher-order nonlinear differential equations are given in references 4 and 5.

Servomechanisms with Two Nonlinear Controllers

Consider the servomechanism with two nonlinear controllers, as shown in Fig. 5. In the forward link, the nonlinear controller is designed by G_{N1}. The transfer function for the principal member to be controlled is given as

$$G_1(s) = \frac{1}{s(Ts+1)} \tag{26}$$

This has the same form as $G(s)$ given in equation 8. In the feedback link let there be another controller, designated by G_{N2}.

To simplify the discussion, let the nonlinear characteristics of G_{N1} and G_{N2} be given as follows:

$$e'' + g_{N1}(e', e) = m_1(t) \tag{27}$$

$$c'' + g_{N2}(c', c) = b(t) \tag{28}$$

where $m_1(t)$ represents the principal control effort and is equal to the input to the principal linear network G_1.

Let $M_1(s)$ be the Laplace transform of $m_1(t)$. Then

Fig. 2. Feedback system with $G_1(s) = \frac{1}{Ts+1}$ and $G_2(s) = \frac{1}{s}$

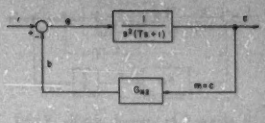


Fig. 3. Feedback system with $G_1(s) = \frac{1}{s^2(Ts+1)}$

$$M_1(s) = \frac{1}{G_1(s)} X(s) = s(Ts+1)C(s) \tag{29}$$

Inversion yields

$$m_1(t) = Tc'' + c' \tag{30}$$

Combining equations 27 and 30 gives

$$e'' + g_{N1}(e', e) = Tc'' + c' \tag{31}$$

Since $b(t) = r(t) - e(t)$, equation 28 becomes

$$c'' + g_{N2}(c', c) = r - e \tag{32}$$

Equations 31 and 32 are two simultaneous nonlinear differential equations of the second order. The solution of similar set of equations is given in reference 2.

Since $c'' = dc'/dt = (dc'/dc)(dc/dt) = c'(dc'/dc)$, equation 32 can be rewritten as

$$c'(dc'/dc) = c'' = r - e - g_{N2}(c', c) \tag{33}$$

Therefrom, the phase-plane equation is

$$\frac{dc'}{dc} = \frac{c''}{c'} = \frac{r - e - g_{N2}(c', c)}{c'} \tag{34}$$

Similarly, another phase-plane equation, relating the rate of actuating signal e' and the signal e, can be obtained as

$$\frac{de'}{de} = \frac{e''}{e'} = \frac{Tc'' + c' - g_{N1}(e', e)}{e'} \tag{35}$$

where

$$c'' = r - e - g_{N2}(c', c) \tag{36}$$

A General System with Two Nonlinear Controllers

Let a linear network be in the forward link and a nonlinear controller in the same link. Let a linear network be in the feedback link and an auxiliary nonlinear controller in the feedback link, as shown in Fig. 6. The following procedure of analyzing the whole system is applicable:

1. Write the Laplace transform equations:

$$G_1(s)M_1(s) = C(s)$$

$$G_N(s)C(s) = M(s)$$

$$R(s) - E(s) = B(s)$$

2. Perform the inversions

$$m_1(t) = \mathcal{L}^{-1}\frac{C(s)}{G_1(s)} = \mathcal{L}^{-1}\frac{M(s)}{G_1(s)G_N(s)}$$

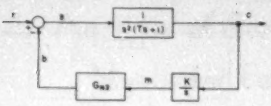


Fig. 4. Feedback system with $G_1(s) = \frac{1}{s^2(Ts+1)}$ and $G_2(s) = \frac{1}{s}$

$$m(t) = \mathcal{L}^{-1}\, G_2(s)C(s) = \mathcal{L}^{-1}\, M(s)$$

3. Write the nonlinear functions characterizing the nonlinear controllers G_{N1} and G_{N2}:

$$g_{n1}(e'',\ e',\ e) = m_1(t)$$

$$g_{n2}(m'',\ m',\ m) = b(i) = r(i) - e(i)$$

4. Combine the results of steps 2 and 3 to get the following simultaneous nonlinear differential equations:

$$g_{n1}(e'',\ e',\ e) = a_n m^{(n)} + \ldots\ldots + a_2 m'' + a_1 m' + a_0 m$$

$$g_{n2}(m'',\ m',\ m) = r - e$$

The following example illustrates the general procedure outlined in the foregoing. Let $G_1(s) = 1/s(Ts+1)$ and let $G_2(s) = 1/s$. Then

$$1/G_1(s)G_2(s) = s^2(Ts+1) \tag{37}$$

The principal control effort $m_1(t)$ is related to the feedback control effort $m(t)$ by

$$m_1(t) = Tm''' + m'' \tag{38}$$

Assuming the same nonlinearities as given in equations 27 and 28 gives

$$e'' + g_{n1}(e',\ e) = Tm''' + m'' \tag{39}$$

$$m'' + g_{n2}(m',\ m) = r - e \tag{40}$$

In order to solve these two nonlinear differential equations simultaneously, m''' can be replaced by terms obtained from the following equation

$$m''' + g_{n2}'(m',\ m) = r' - e' \tag{41}$$

where $g_{n2}' = dg_{n2}/dt$ is a function which contains m'' and m' and possibly also m. Thus equation 39 becomes

$$e'' = T[(r' - e') - g_{n2}'] + [(r - e) - g_{n2}] - g_{n1}(e',\ e) \tag{42}$$

which involves the second derivatives of both variables e and m. The phase-plane equation for equation 40 is similar to equation 34.

$$\frac{dm'}{dm} = \frac{m''}{m'} = \frac{r - e - g_{n2}(m',\ m)}{m'} \tag{43}$$

Similarly, the phase-plane equation obtained from equation 39 is

$$\frac{de'}{de} = \frac{Tm''' + m'' - g_{n1}(e',\ e)}{e'} \tag{44}$$

where m''' and m'' can be replaced by terms obtained from equations 41 and 40.

Equations 43 and 44 are two phase-plane equations that can be solved simultaneously, by the method given in reference 2.

The 4-dimensional space involved has as its co-ordinates m, m', e, and e'. The two equations represent two trajectories in the $m'-m$ plane and the $e'-e$ plane. The second derivative of m is given by

$$m'' = r - e - g_{n2}(m',\ m) \tag{45}$$

Third derivative of m is given by

$$m''' = r' - e' - g_{n2}' \tag{46}$$

where g_{n2}' is a function of m'', m', and m. Finally, the second derivative of the actuating signal is given by

$$e'' = Tm''' + m'' - g_{n1}(e',\ e) \tag{47}$$

Thus four variables m, m', e, and e' are sufficient to represent a complete system with two nonlinear control functions. For higher-order systems, a third phase-plane equation may be introduced as

$$\frac{dm''}{dm'} = \frac{m'''}{m''} \tag{48}$$

where m''' can be expressed in terms of m'', m', m, if a third-order nonlinear differential equation is involved.

Conclusions

There have been many contributions for utilizing nonlinear components to stabilize a feedback system and for developing synthesis techniques for designing systems with nonlinear controllers. This paper gives a procedure for analyzing any servomechanism when such synthesis techniques yield suggested alternative designs. For systems with nonlinear feedback control, one may start with the form of desired output, and then work backwards to get the desired actuating signal function. The difference between the given input and the actuating signal gives the desired feedback signal. If there is no linear network in the feedback link, the problem reduces to finding the

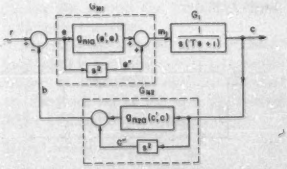


Fig. 5. Feedback system with two nonlinear controllers

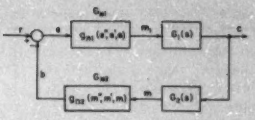


Fig. 6. Feedback system with two linear networks and two nonlinear controllers

correct nonlinear function $g_{n3}(c'',\ c',\ c) = b = r - e$. If there is a linear network in the feedback path, it is necessary to find the inverse transform of $G_2(s)C(s) = m(s)$, or the feedback control effort m. Then the problem is to find the nonlinear control function $g_{n2}(m'',\ m',\ m) = b = r - e$.

In case there are two nonlinear controllers (one in the forward link and one in the feedback link), the principal control effort $m_1(t)$ can be found by taking the inverse transform of $M_1(s) = C(s)/G_1(s)$. Then it is necessary to find the principal nonlinear control function $g_{n1}\ (e'',\ e',\ e) = m_1(t)$ and the auxiliary nonlinear control function $g_{n2}\ (m'',\ m',\ m) = r - e$ simultaneously. The techniques for analyzing such a general system involve the simltaneous solutions of two first-order nonlinear differential equations. During the process of solution, one sees the relation between the actuating signal and the rate of actuating signal, and also the relation between the feedback control effort and the time derivative of this effort. The two trajectories thus give insight as to the stability and transient behavior of the system under control.

Appendix

Two examples given here are based on references 15 and 10.

Example I

The block diagram of the system is shown in Fig. 7 (see Fig. 1, reference 15).

The relations between the different quantities can be expressed by the following equations:

$$r - c = e \tag{49}$$

$$e + m_1 - m_2 = m \tag{50}$$

$$m_1 = Kec' = K(r - c)c' \tag{51}$$

$$m_2 = K_2 c' \tag{52}$$

$$m = Tc'' + c' \tag{53}$$

The last two equations are obtained from the inverse transforms of the Laplace transform equations

$$M(s)G_1(s) = C(s) \tag{54}$$

$$C(s)G_2(s) = M_2(s) \tag{55}$$

where

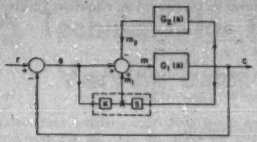


Fig. 7. Block diagram of Example I

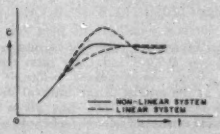


Fig. 8. Response curve of Example I

$$G_1(s)=\frac{1}{s(Ts+1)} \tag{56}$$

$$G_2(s)=K_2s \tag{57}$$

From equations 49 to 53, the following second-order nonlinear differential equation is obtained:

$$Tc''+(1+K_2)c'-K(r-c)c'+c=r \tag{58}$$

Let

$$A=(1+K_2)$$

and

$$B=A-Kr$$

Equation 58 becomes

$$Tc''+(B+Kc)c'+c=r \tag{59}$$

If a unit step $u(t)$ is used as the input $r(t)$, B can be considered as a constant. The design of the positional servomechanism aims to have the output member following the instantaneous position of the input member. The experimental response curves are shown in Fig. 8 which is taken from Fig. 5 of reference 15. (The calculated result was given in Fig. 3 of reference 15.)

Example II

The block diagram of the system is shown in Fig. 9 (see Fig. 3-2, reference 10). A relay which limits the value of m to F is not shown in the simplified diagram. The diode network can be designated as $g_{s1}(e)$.

The relations between the different quantities can be expressed by the following equations:

$$r-c=e \tag{60}$$

$$m_1=g_{s1}(e) \tag{61}$$

$$e'-m_1=m \text{ (} m \text{ is limited by } \pm F_{max}\text{)} \tag{62}$$

$$m=Tc'''+c'' \tag{63}$$

The last equation is obtained from the inversion of the Laplace transform equation

$$M(s)G_1(s)=C(s) \tag{64}$$

where

$$G_1(s)=\frac{1}{s^2(Ts+1)} \tag{65}$$

From equations 60 to 63, the following third-order nonlinear differential equation is obtained.

$$Tc'''+c''+c'=r'-g_{s1}(e) \tag{66}$$

Substituting $c'''=r'''-e'''$, $c''=r''-e''$, and $c'=r'-e'$ gives

$$Te'''+e''+[e'-g_{s1}(e)]=Tr'''+r'' \tag{67}$$

Since the input function $r(t)$ is known, the right-hand member of equation 67 is also known. The left-hand member has three terms. The third term (in square brackets) represents the control effort m which is limited by the values $\pm F_{max}$. This is a third-order nonlinear differential equation in e, which can be solved by the method outlined in reference 4.

The trajectory showing the relation between e' and e for the case of a ramp input was shown in Fig. 3-4 of reference 10. When e is known, the output c is equal to $r-e$. The output or response curves are shown in solid lines in Fig. 10, while the ramp inputs are shown in dotted lines (see Fig. 3-5 of reference 10).

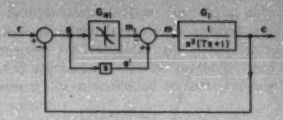


Fig. 9. Block diagram of Example II

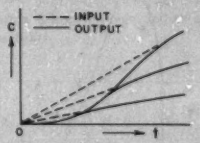


Fig. 10. Typical ramp responses of Example II

References

1. Analysis of Nonlinear Coupled Circuits, Y. H. Ku. *AIEE Transactions*, vol. 73, pt. I, 1954 (Jan. 1955 section), pp. 626–31.

2. Analysis of Nonlinear Coupled Circuits—II, Y. H. Ku. *Ibid.*, vol. 74, Sept. 1955, pp. 439–43.

3. Acceleration Plane Method for Analysis of a Circuit with Nonlinear Inductance and Nonlinear Capacitance, Y. H. Ku. *Ibid.*, vol. 73, pt. I, 1954 (Jan. 1955 section), pp. 619–26.

4. A Method for Solving Third and Higher Order Nonlinear Differential Equations, Y. H. Ku. *Journal*, Franklin Institute, Philadelphia, Pa., vol. 256, no. 3, Sept. 1953, pp. 229–44.

5. Analysis of Nonlinear Systems with More Than One Degree of Freedom by Means of Space Trajectories, Y. H. Ku. *Ibid.*, vol. 259, no. 2, Feb. 1955, pp. 115–31.

6. Nonlinear Techniques for Improving Servo Performance, D. McDonald. *Proceedings*, National Electronics Conference, Chicago, Ill., vol. 5, 1950, pp. 400–21.

7. A Frequency Response Method of Analyzing and Synthesizing Contactor Servomechanisms, Ralph J. Kochenburger. *AIEE Transactions*, vol. 69, pt. I, 1950, pp. 270–84.

8. Sinusoidal Analysis of Feedback-Control Systems Containing Nonlinear Elements, E. Calvin Johnson. *Ibid.*, vol. 71, pt. II, July 1952, pp. 169–81.

9. Phase-Plane Analysis of Automatic Control Systems with Nonlinear Gain Elements, R. E. Kalman. *Ibid.*, vol. 73, pt. II, 1954 (Jan. 1955 section), pp. 383–90.

10. Nonlinear Control of a Saturating Third-Order Servomechanism, J. L. Preston. *Technical Memorandum 6897-TM-14*, Servomechanism Laboratory, Massachusetts Institute of Technology, Cambridge, Mass., May 1954.

11. Analysis and Design Principles of Second and Higher Order Saturating Servomechanisms, R. E. Kalman. *AIEE Transactions*, vol. 74, pt. II, Nov. 1955, pp. 294–310.

12. Quasi-Linearization Techniques for Transient Study of Nonlinear Feedback Control Systems, Kan Chen. *Ibid.*, (Jan. 1956 section), pp. 354–65.

13. A Method of Evaluating Nonlinear Servomechanisms, M. V. Mathews. *Ibid.*, May 1955, pp. 114–23.

14. A Generalized Method for Determining the Closed-Loop Frequency Response of Nonlinear Systems, Luther T. Prince, Jr. *Ibid.*, vol. 73, pt. II, Sept. 1954, pp. 217–24.

15. The Use of Nonlinear Feedback to Improve the Transient Response of a Servomechanism, J. B. Lewis. *Ibid.*, vol. 71, pt. II, 1952 (Jan. 1953 section), pp. 449–53.

Discussion

Thomas J. Higgins (University of Wisconsin, Madison, Wis.): I find this a lucidly written and topically interesting paper. It will direct the attention of servo engineers to the usefulness in practice of Professor Ku's very unique graphical methods for the solution of higher-order nonlinear differential equations. It would be of interest actually to have the details of solution by the author's procedures in references 1 to 5 for one of the illustrative examples of the Appendix. Could he provide this in his closure?

Y. H. Ku: I wish to thank Professor Higgins for his comments. Example I in the Appendix was given as a class problem in a graduate course on nonlinear circuit analysis offered in 1955–56. Using a unit step as the input r, equation 59 becomes

$$Tc''+Bc'+Kcc'+c=1 \tag{68}$$

where $B=A-K$, a constant. Let $\tau=\omega_0 t$, where $\omega_0^2=1/T$. Equation 68 can then be normalized to

$$\frac{d^2c}{d\tau^2}+\alpha\frac{dc}{d\tau}+\beta\frac{d(c^2)}{d\tau}+c=1 \tag{69}$$

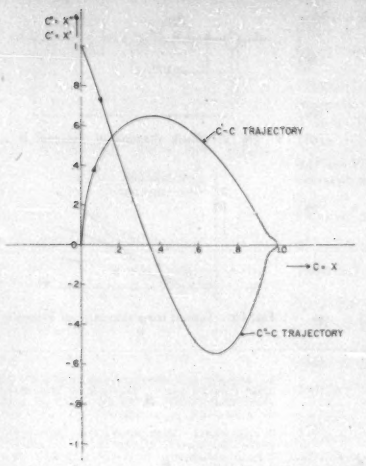


Fig. 11. Trajectories for Example I of the Appendix

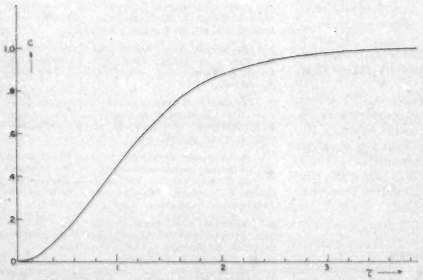


Fig. 12. Transient response curve for Example I of the Appendix

where $\alpha=\omega_0 B$ and $\beta=K\omega_0/2$. The initial conditions are: at $\tau=0$, $c=0$ and $dc/d\tau=0$.

C. J. Lin, formerly of this class, and now in the Physics Department, University of Pennsylvania, used Hsia's method[1] to solve equation 69 and considered five possible cases:

1. $\alpha=0.10$, $\beta=0.45$
2. $\alpha=0.10$, $\beta=0.90$
3. $\alpha=0.10$, $\beta=1.00$
4. $\alpha=0.20$, $\beta=0.90$
5. $\alpha=0.10$, $\beta=1.30$

Case 5 was found to give the right answer to the desired performance.

Morten Kanefsky, a senior student at the Moore School, followed my "acceleration plane method"[2,3] and worked out the solution of equation 69 for $\alpha=0.10$ and $\beta=1.30$ corresponding to Lin's case 5.

Let $c=x$, $dc/d\tau=x'=v$ and $d^2c/d\tau^2=x''=w$. Equation 69 can be rewritten as

$$w+(\alpha+2\beta x)v+x=1 \tag{70}$$

Putting in numerical values of α and β gives

$$w+(0.1+2.6x)v+x=1 \tag{71}$$

Let $\Delta x=0.01$. Since the initial value of x is 0, the value of x at the end of the small interval is 0.01. Substituting $x=0.01$ into equation 71 gives

$$w=(1-0.01)-(0.1+0.026)v=0.99-0.126v \tag{72}$$

Now the phase-plane equation is

$$\frac{dv}{dx}=\frac{w}{v}=\frac{(1-x)-(0.1+2.6x)v}{v} \tag{73}$$

The difference equation for the small interval is

$$\frac{\Delta v}{\Delta x}=\frac{0.99-0.126v}{v} \tag{74}$$

Since the initial value of v is given as 0, the increment Δv is equal to the value of v at the end of the interval. Denoting the new values of w, v, and x at the end of the interval by w_1, v_1, and x_1, equation 74 gives

$$\frac{v_1}{0.01}=\frac{w_1}{v_1}=\frac{0.99-0.126v_1}{v_1} \tag{75}$$

Solving the quadratic equation

$$100v_1^2+0.126v_1-0.99=0 \tag{76}$$

gives $v_1=0.0988$, the positive root being chosen as both Δx and Δt are positive. The slope of the $v-x$ trajectory at the end of the interval is

$$\frac{dv}{dx}=9.88 \tag{77}$$

The corresponding values of w_1 and v_1 are 0.9775 and 0.0988 and their ratio checks with $\Delta v/\Delta x$. As discussed before,[2,4] the values at the middle of the interval should be chosen. The author has given the numerical results of Mr. Kanefsky to show that if the chosen value of Δx is small, a good check can be obtained about the slope at the end of that interval. Note that the initial value of w, the second derivative, is 1, according to equation 70, for $v=0$ and $x=0$. The initial slope is $(dv/dx)=1/0$, which is infinite. The slope is reduced to about 10 at the end of the chosen interval, according to equation 77.

In further steps, Mr. Kanefsky used the equivelocity lines (or curves in general) in the $w-x$ plane to construct the $v-x$ trajectory. The $v-x$ trajectory is shown as c'-c trajectory in Fig. 11. The $w-x$ trajectory is shown as c''-c trajectory in the same figure. The equivelocity lines (or curves) are omitted from the original plot. The transient response curve is shown in Fig. 12, with τ as the abscissa. The instantaneous curve checks numerically with Lin's results and agrees with the solid-line curve shown in Fig. 8 of the paper. The trajectories and the instantaneous curve are further checked by the results obtained from a differential analyzer by Kanefsky and C. F. Chen.

It may be of interest to mention here that in a paper presented at the international conference on "Fachtagung Regelungstechnik" at Heidelberg, Germany, September 1956, by B. N. Naumov of Moscow, four of my papers were mentioned in connection with the phase-plane method.[2-4] This information was confirmed by Harold Chestnut, chairman of the AIEE session on feedback systems at the Fall General Meeting during which this paper was presented.

References

1. A Graphical Analysis for Nonlinear Systems, Pei-Su Hsia. *Proceedings*, Institution of Electrical Engineers, London, England, vol. 99, pt. II, 1952, pp. 125–34.

2. Nonlinear Analysis of Electromechanical Problems, Y. H. Ku. *Journal*, Franklin Institute, Philadelphia, Pa., vol. 255, 1953, pp. 9–31.

3. An Acceleration Plane Method for Nonlinear Oscillations, Y. H. Ku. *Proceedings*, Symposium on Nonlinear Circuit Analysis, Polytechnic Institute of Brooklyn, Brooklyn, N. Y., 1953, pp. 129–52.

4. Analysis of Multi-Loop Nonlinear Systems, Y. H. Ku. *Transactions*, Professional Group on Circuit Theory-1, Institute of Radio Engineers, New York, N. Y., no. 4, 1954, pp. 6–12.

5. See reference 4 of the paper.

A Technique for the Analysis of Time-Varying Sampled-Data Systems

BERNARD FRIEDLAND
NONMEMBER AIEE

MUCH OF THE LITERATURE in the field of feedback control has been concerned with linear, time-invariant systems. The labor involved in obtaining solutions to even simple problems in time-varying systems may well have discouraged research in this area. Nevertheless, many practical systems unavoidably contain time-varying components, and methods of dealing with such systems are desirable.

In this paper a method for analyzing sampled-data systems containing time-varying components is discussed. Each component of such a system is presumed to have an input comprising a train of amplitude-modulated pulses. For the purposes of numerical analysis these systems are easier to deal with than continuous systems.[1] The fact that in sampled-data systems one deals with sequences of numbers (the amplitudes of the pulses circulating through the system) rather than with continuous time functions makes this method of analysis possible.

A typical component of a sampled-data system is shown in Fig. 1. It consists of a linear, time-varying element H preceded and followed by sampling switches. These switches convert a continuous time function into a train of pulses whose amplitudes are determined by the value of the continuous time function at the sampling instant, i.e.,

$$x^*(t) = \sum_{k=0}^{\infty} x(kT)\,p(t-kT) \qquad (1)$$

where $p(t-kT)$ is a pulse occurring at $t=kT$. The continuous output of the component $y(t)$ is given by the discrete form of the superposition equation

$$y(t) = \sum_{k=0}^{\infty} x(kT)\,h(t,\,kT) \qquad (2)$$

where $h(t,\tau)$ is the response of H to $p(t-\tau)$, a pulse of unit amplitude supplied by the switch at $t=\tau$. At the sampling instants the output is

$$y(nT) = \sum_{k=0}^{\infty} x(kT)\,h(nT,\,kT) \qquad (3)$$

The sampled output from the component is

$$y^*(t) = \sum_{n=0}^{\infty} y(nT)\,p(t-nT) \qquad (4)$$

It is noted that $p(t-\tau)$ is a narrow pulse occurring at $t=\tau$ as a result of the sampling switch closing at this instant when the continuous function is of unit amplitude. The exact shape of this pulse depends on the characteristics of the switch. It is frequently convenient, but not essential, to assume that the switch produces mathematical impulses (delta functions). In this case $p(t-\tau)=\delta(t-\tau)$ and $h(t,\tau)$ is the impulsive response of H.[2] It is noted further that if H is a fixed (time-invariant) component then $h(t,\tau)=h(t-\tau)$.

It is assumed in this paper that all components of a control system are separated from each other by identical sampling switches which are synchronized to open and close together. In this case the sampling interval T is constant for all components, with inputs and outputs consisting of trains of amplitude-modulated, but otherwise identical, pulses, and the present concern is with the input-output relation for the amplitudes of the pulses

$$y(n) = \sum_{k=0}^{\infty} h(n,\,k)x(k) = \sum_{k=0}^{n} h(n,\,k)x(k) \qquad (5)$$

In this equation the upper limit is changed from ∞ to n because the response of a physical system at rest prior to the application of an input is zero. Hence

$$h(n,\,k)=0 \text{ for } k>n$$

Paper 56-962, recommended by the AIEE Feedback Control Systems Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Fall General Meeting, Chicago, Ill., October 1–5, 1956. Manuscript submitted October 17, 1955; made available for printing July 24, 1956.

BERNARD FRIEDLAND is with Columbia University, New York, N. Y.

The author gratefully acknowledges the advice and support of Prof. J. R. Ragazzini under whose guidance this investigation was conducted; and also wishes to thank Prof. L. A. Zadeh and other members of the Department of Electrical Engineering of Columbia University for a number of helpful discussions. This research was supported by the U.S. Air Force under Contract No. *AF 18(600)-677*, and monitored by the Office of Scientific Research, Air Research and Development Command.

Sampled-data systems containing only fixed elements have been studied extensively by transform (generating function) methods.[3-6] These methods have recently been extended to time-varying systems.[7]

Algebra of Time-Varying Systems

Equation 5 may be regarded as a matrix equation

$$\mathbf{y}=\mathbf{H}\mathbf{x} \qquad (6)$$

where

$$\mathbf{y}=\begin{bmatrix} y(0) \\ y(1) \\ \cdot \\ \cdot \\ \cdot \\ y(n) \end{bmatrix} = \text{col } [y(0),\, y(1),\, \ldots\, y(n)] = \text{output signal vector}$$

$$\mathbf{x}=\begin{bmatrix} x(0) \\ x(1) \\ \cdot \\ \cdot \\ \cdot \\ x(n) \end{bmatrix} = \text{col } [x(0),\, x(1),\, \ldots\, x(n)] = \text{input signal vector}$$

and

$$\mathbf{H}=\begin{bmatrix} h(0,0) & 0 & 0 & \ldots \\ h(1,0) & h(1,1) & 0 & \ldots \\ h(2,0) & h(2,1) & h(2,2) & \ldots \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \\ h(n,0) & h(n,1) & h(n,2) & \ldots \end{bmatrix} = \text{transmission matrix for } H$$

The transmission matrices for several components commonly encountered in feedback systems are given in Table I. In the analysis which follows it is assumed that the transmission matrices for all the components in the system are known. For fixed components the transmission matrix may be obtained from the z-transform[8] of the transmission function $H(s)$, by expanding the latter in a Maclaurin series in z^{-1} and writing the coefficients of the expansion along the rows of the matrix, starting with the principal diagonal and working to the left. For example, the z-transform of a discrete integrator is

$$H(z)=1/(1-z^{-1})=1+z^{-1}+z^{-2}+z^{-3}+\ldots$$

from which the transmission matrix in Table I is obtained. A large proportion of time-varying systems may be regarded as comprising only fixed components and amplifiers with time-varying gain. The transmission matrices of all components of such systems are readily available. If a system contains time-varying components of the general type, the transmission matrices of these components

Fig. 1. Typical component of time-varying sampled-data system

must be obtained, either by solving the difference equation relating the output and input or, experimentally, by applying sampling pulses to the component and observing the output. Since only the amplitudes of the output as a series of numbers is required, rather than the continuous time function, the utility of the experimental method should not be overlooked.

A time-varying sampled-data system consists of a number of components interconnected through sampling switches. The objective is to determine the transmission matrix for such a system, given the transmission matrices of all the components in the system.

A parallel combination of two components is shown in Fig. 2(A). By the principle of superposition it is clear that the transmission matrix for the combination is simply the sum of the transmission matrices of the components. A series combination of two components is shown in Fig. 2(B). Let the input signal to the first component H be x. Then the output of this component is given by

$$\mathbf{y}=\mathbf{Hx}$$

But y is the input to the second component G. Hence the output from this component is given by

$$\mathbf{w}=\mathbf{Gy}=\mathbf{GHx}$$

Thus the transmission matrix for the series combination of H followed by G is given by the matrix product **GH**. It is important to note that since matrix multiplication is not commutative, the series combination of H followed by G is not equal, in general, to the series combination of G followed by H. (If G and H are both fixed, **GH** does equal **HG**.)

Using the rules for the combination of components in a system developed above enables one to determine the transmission matrix for a feedback system of arbitrary complexity. The transmission matrix for the unity feedback system shown in Fig. 3 is derived as an example.

$$\mathbf{e}=\mathbf{r}-\mathbf{c}$$

$$\mathbf{c}=\mathbf{GDe}$$

Thus

$$\mathbf{e}=\mathbf{r}-\mathbf{GDe}$$

Adding **GDe** to both sides yields

$$\mathbf{e}+\mathbf{GDe}=(\mathbf{I}+\mathbf{GD})\mathbf{e}=\mathbf{r}$$

where **I** is the identity matrix. Premultiplying both sides of this equation by $(\mathbf{I}+\mathbf{GD})^{-1}$ yields

$$\mathbf{e}=(\mathbf{I}+\mathbf{GD})^{-1}\mathbf{r}$$

Hence

$$\mathbf{c}=\mathbf{GD}(\mathbf{I}+\mathbf{GD})^{-1}\mathbf{r}$$

Thus the transmission matrix of the system, defined by the relation $\mathbf{c}=\mathbf{Kr}$, is given by

$$\mathbf{K}=\mathbf{GD}(\mathbf{I}+\mathbf{GD})^{-1} \tag{7}$$

It can be shown that $\mathbf{F}(\mathbf{I}+\mathbf{F})^{-1}=(\mathbf{I}+\mathbf{F})^{-1}\mathbf{F}$. Hence the following may be written unambiguously

$$\mathbf{K}=\mathbf{GD}/(\mathbf{I}+\mathbf{GD}) \tag{8}$$

It can also be shown that

$$\mathbf{K}=\mathbf{I}-(\mathbf{I}+\mathbf{GD})^{-1} \tag{9}$$

The latter form is frequently convenient for computation.

Inversion of Triangular Matrices

The numerical analysis of time-varying feedback systems involves the evaluation of matrices of the form $(\mathbf{I}+\mathbf{H})$. Since physical systems have uniquely determined finite outputs, these matrices are always nonsingular.

The order of the matrix to be inverted depends on the number of output pulses required. If more than three or four output pulses are needed, inversion of $\mathbf{I}+\mathbf{H}$ by the customary method is an arduous arithmetical task. Fortunately, however, for physical systems, $\mathbf{I}+\mathbf{H}$ will always be a "triangular" matrix, and can be inverted with relatively little effort by the well-known method to be outlined.

Consider the equation

$$\mathbf{Mx}=\mathbf{y} \tag{10}$$

where **x** and **y** are column vectors and H is a triangular matrix. The inverse of **M** is given then by the equation

$$\mathbf{x}=\mathbf{M}^{-1}\mathbf{y} \tag{11}$$

Now, equation 10 implies the following set of simultaneous equations

$$m_{00}x_0 = y_0 \tag{12}$$

$$m_{10}x_0+m_{11}x_1 = y_1 \tag{12A}$$

$$m_{20}x_0+m_{21}x_1+m_{22}x_2 = y_2 \tag{12B}$$

$$\vdots$$

$$m_{n0}x_0+m_{n1}x_1+m_{n2}\ x_2+\ldots+m_{nn}x_n = y_n(s)$$

This set of equations can be solved successively for $x_0, x_1, \ldots x_n$ in terms of $y_0, y_1, \ldots y_n$, by solving equation 12 for x, substituting this solution into equation 12(A), and solving for x_1. Substituting x_0 and x_1 into equation 12(B) solves for x_2. This process is repeated successively n times and results in

$$\begin{aligned} x_0 &= u_{00}y_0 \\ x_1 &= u_{10}y_0+u_{11}y_1 \\ x_2 &= u_{20}y_0+u_{21}y_1+u_{22}y_2 \\ &\vdots \\ x_n &= u_{n0}y_0+u_{n1}y_1+\ldots+u_{nn}y_n \end{aligned} \tag{13}$$

But the set of solutions of equation 13 may be written as

Table I. Transmission Matrices for Common Components

Components	Matrices
Fixed component $h(n,k)=h(n-k)$	$\begin{bmatrix} h(0) & 0 & 0 & 0 & \cdots \\ h(1) & h(0) & 0 & 0 & \cdots \\ h(2) & h(1) & h(0) & 0 & \cdots \\ h(3) & h(2) & h(1) & h(0) & \cdots \\ & & \vdots & & \end{bmatrix}$
Amplifier with variable gain	$\begin{bmatrix} h(0) & 0 & 0 & 0 & \cdots \\ 0 & h(1) & 0 & 0 & \cdots \\ 0 & 0 & h(2) & 0 & \cdots \\ 0 & 0 & 0 & h(3) & \cdots \\ & & \vdots & & \end{bmatrix}$
Discrete integrator $h(n,k)=1,\ k\leq n$ $=0,\ k>n$	$\begin{bmatrix} 1 & 0 & 0 & 0 & \cdots \\ 1 & 1 & 0 & 0 & \cdots \\ 1 & 1 & 1 & 0 & \cdots \\ 1 & 1 & 1 & 1 & \cdots \\ & & \vdots & & \end{bmatrix}$
Discrete differentiator $h(n,k)=1,\ k=n$ $=-1,\ k=n-1$ $=0,\ k>n$ $=0,\ k<n-1$	$\begin{bmatrix} 1 & 0 & 0 & 0 & \cdots \\ -1 & 1 & 0 & 0 & \cdots \\ 0 & -1 & 1 & 0 & \cdots \\ 0 & 0 & -1 & 1 & \cdots \\ & & \vdots & & \end{bmatrix}$

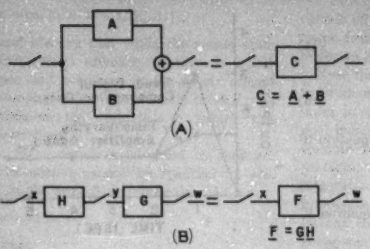


Fig. 2 (left). Parallel and series combination of components

$$x = Uy \qquad (14)$$

Comparing equation 14 with equation 11, it is observed that $\mathbf{U}=\mathbf{M}^{-1}$, and that the process of successive solution of algebraic equations each in one variable, has accomplished the inversion of a triangular matrix.

This result is not surprising, since it simply implies that the first n pulses in the output of a system, depend on, at most, only the first n input pulses.

Numerical Example

A numerical example of a unity feedback sampled-data system, shown in Fig. 3, will be solved by the methods developed in the foregoing. The sampling switches will be assumed to produce very narrow pulses closely approximating impulses, and have a period of 1 second. D is an amplifier with time-varying gain given by

$$d(t) = 1 + \epsilon^{-t}$$

Thus

$$\mathbf{D} = \begin{bmatrix} 1+1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 1+\epsilon^{-1} & 0 & 0 & 0 & \cdots \\ 0 & 0 & 1+\epsilon^{-2} & 0 & 0 & \cdots \\ 0 & 0 & 0 & 1+\epsilon^{-3} & 0 & \cdots \\ 0 & 0 & 0 & 0 & 1+\epsilon^{-4} & \cdots \\ & & & \vdots & & \end{bmatrix} = \begin{bmatrix} 2.000 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 1.368 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 1.135 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 1.050 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 1.018 & \cdots \\ & & & \vdots & & \end{bmatrix}$$

Component G is a fixed component consisting of a zero-order hold followed by an integrator, as shown in Fig. 4(A). From Fig. 4(B) it is seen that

$$\mathbf{G} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & \cdots \\ 1 & 0 & 0 & 0 & 0 & \cdots \\ 1 & 1 & 0 & 0 & 0 & \cdots \\ 1 & 1 & 1 & 0 & 0 & \cdots \\ 1 & 1 & 1 & 1 & 0 & \cdots \\ & & \vdots & & & \end{bmatrix}$$

Forming the matrix product $\mathbf{GD}$ yields

$$\mathbf{GD} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & \cdots \\ 2.000 & 0 & 0 & 0 & 0 & \cdots \\ 2.000 & 1.368 & 0 & 0 & 0 & \cdots \\ 2.000 & 1.368 & 1.135 & 0 & 0 & \cdots \\ 2.000 & 1.368 & 1.135 & 1.050 & 0 & \cdots \\ & & & \vdots & & \end{bmatrix}$$

and

$$\mathbf{I}+\mathbf{GD} = \begin{bmatrix} 1.000 & 0 & 0 & 0 & 0 & \cdots \\ 2.000 & 1.000 & 0 & 0 & 0 & \cdots \\ 2.000 & 1.368 & 1.000 & 0 & 0 & \cdots \\ 2.000 & 1.368 & 1.135 & 1.000 & 0 & \cdots \\ 2.000 & 1.368 & 1.135 & 1.050 & 1.000 & \cdots \\ & & & \vdots & & \end{bmatrix}$$

This matrix is inverted by the method of the previous section. The result is

$$(\mathbf{I}+\mathbf{GD})^{-1} = \begin{bmatrix} 1.000 & 0 & 0 & 0 & 0 & \cdots \\ -2.000 & 1.000 & 0 & 0 & 0 & \cdots \\ 0.736 & -1.368 & 1.000 & 0 & 0 & \cdots \\ -0.099 & 0.194 & -1.135 & 1.000 & 0 & \cdots \\ 0.005 & -0.010 & 0.057 & -1.050 & 1.000 & \cdots \\ & & \vdots & & & \end{bmatrix}$$

Thus, from equation 9 the transmission matrix for the system is

$$\mathbf{K} = \mathbf{I} - (\mathbf{I}+\mathbf{GD})^{-1} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & \cdots \\ 2.000 & 0 & 0 & 0 & 0 & \cdots \\ -0.736 & 1.368 & 0 & 0 & 0 & \cdots \\ 0.099 & -0.194 & 1.135 & 0 & 0 & \cdots \\ -0.005 & 0.010 & -0.057 & 1.050 & 0 & \cdots \\ & & \vdots & & & \end{bmatrix}$$

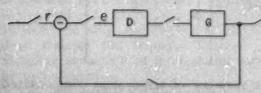


Fig. 3. Unity feedback control system. Switches in input line and in feedback line may be omitted without affecting operation of system. Switch in output line drawn to indicate that output at sampling instants only is considered

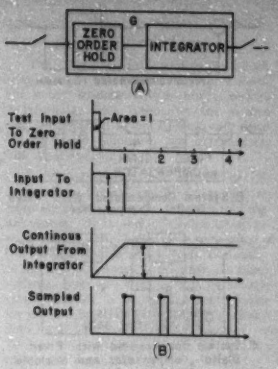


Fig. 4. Component G in numerical example

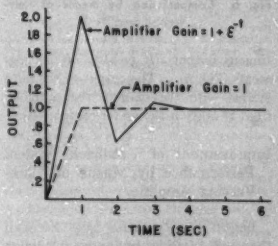


Fig. 5. Response of system in example to unit step input

If the input $r(t)$ to the system is a unit step, then

$$\mathbf{r} = \text{col}\,[1, 1, 1, 1, 1, \ldots]$$

and the output pulse train given by

$$\mathbf{c} = \mathbf{Kr}$$

is, numerically

$$\mathbf{c} = \text{col}\,[0, 2.000, 0.632, 1.050, 0.998, \ldots]$$

Since the second component of G is an integrator, the physical input to which is a staircaselike time function, the con-

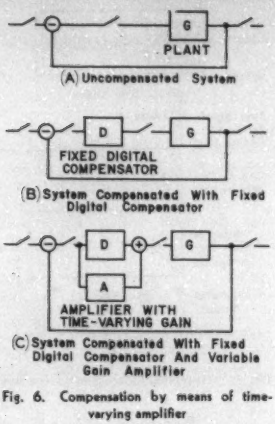


Fig. 6. Compensation by means of time-varying amplifier

tinuous output $c(t)$ consists of line segments (Fig. 5). The output of the same system with the gain of the amplifier fixed at unity is shown for comparison.

Improvement of Feedback System Performance by Means of Time-Varying Amplifier

Bergen and Ragazzini[3] have discussed the compensation of a sampled-data system such as that of Fig. 6, by means of a fixed digital compensator. The resulting system, using their technique is shown in Fig. 6(B). Component D is a fixed digital compensator.

If the plant G has no poles or zeros in the right half s-plane, it is possible to design a digital compensator so that the response at sampling instants to any input is any arbitrary time function consistent with the available driving power. If, on the other hand, the plant has either poles or zeros in the right half s-plane, there are limitations on the response for which the system may be designed, because in practice it is impossible to cancel such singularities by physical elements. It is usually possible to design the system for a satisfactory steady-state response, but the transient response is frequently unsatisfactory.

By introducing a time-varying amplifier A in parallel with the fixed digital compensator, as shown in Fig. 6(C), it is often possible to improve the transient response. The relation by which A is designed is derived as follows: Let

$\mathbf{r}$ = reference input signal vector
$\mathbf{c}$ = desired output signal vector
$\mathbf{c}'$ = output signal vector of system having only fixed digital compensator

Then

$$(1+GD)\mathbf{c}' = GD\mathbf{r} \tag{15}$$

and

$$(1+GD+GA)\mathbf{c} = (GD+GA)\mathbf{r} \tag{16}$$

Subtracting equation 15 from equation 16 and rearranging terms yields

$$(1+GD)(\mathbf{c}-\mathbf{c}') = GA(\mathbf{r}-\mathbf{c}) \tag{17}$$

Premultiplying both sides of equation 17 by $\mathbf{G}^{-1}$ gives

$$(\mathbf{G}^{-1}+\mathbf{D})(\mathbf{c}-\mathbf{c}') = \mathbf{A}(\mathbf{r}-\mathbf{c}) \tag{18}$$

For convenience, let

$$(\mathbf{G}^{-1}+\mathbf{D})(\mathbf{c}-\mathbf{c}') = \mathbf{q} = \text{col}\,[q(0), q(1), \ldots] \tag{19}$$

and

$$(\mathbf{r}-\mathbf{c}) = \mathbf{e} = \text{col}\,[e(0), e(1), \ldots] \tag{20}$$

Then, from equation 18

$$\mathbf{q} = \mathbf{Ae}$$

Now, if A is to be an amplifier with time-varying gain, then its gain at the kth sampling instant is given by

$$a(k) = q(k)/e(k) \tag{21}$$

It is recognized that $e(k)$ is the system error, i.e., the difference between the desired output and the reference input. In a finite settling-time system the error goes to zero after a finite number, say M, of samples, and gives

$$e(k) = 0, \quad \text{for } k > M$$

Since the gain of the amplifier cannot be made infinite, the desired output must be adjusted so that

$$q(k) = 0, \quad \text{for } k > N \geq M \tag{22}$$

For $N > M$, with $q(k)$ satisfying the restriction of equation 22, $a(k) = 0$. For $N = M$, $a(k)$ is a 0/0 form, and can be set to any convenient value, including zero, which is probably the most convenient.

These results imply that if the desired output can be adjusted to meet the requirements of equation 22, the system can be compensated by means of a time-varying amplifier whose gain is adjusted in accordance with equation 21.

Numerical Example

Consider a plant whose transmission function is

$$G(z) = 1/1-2z^{-1}$$

This plant has a pole outside the unit circle in the z-plane which means it has a pole in the right half of the s-plane. The transmission function of the fixed digital compensator required to stabilize

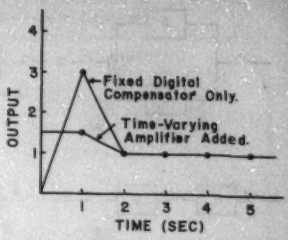


Fig. 7. Response to unit step in example of compensation by means of time-varying amplifier

the plant, according to the theory developed by Bergen and Ragazzini, is

$$D(z) = \frac{1}{G(z)} \frac{K(z)}{1-K(z)} \tag{23}$$

where $K(z)$ is the desired transmission function of the system. Due to the pole in $G(z)$, the following restrictions must be applied to $K(z)$:

1. To stabilize the system with real elements $1-K(z)$ must include a factor $1-2z^{-1}$.
2. To follow a step input with zero steady-state error $1-K(z)$ must include a factor $1-z^{-1}$.

Thus, in the simplest allowable form

$$1-K(z) = (1-2z^{-1})(1-z^{-1}) = 1-3z^{-1}+2z^{-2}$$

and

$$K(z) = 3z^{-1}-2z^{-2}$$

Solving equation 23 for $D(z)$ yields

$$D(z) = (3z^{-1}-2z^{-2})/(1-z^{-1})$$

and

$$D(z)+1/G(z) = \frac{3z^{-1}-2z^{-2}}{1-z^{-1}} + 1-2z^{-1} = \frac{1}{1-z^{-1}} \tag{24}$$

Expanding equation 24 in a Maclaurin series in z^{-1} yields

$$\mathbf{D}+\mathbf{G}^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & \ldots \\ 1 & 1 & 0 & 0 & \ldots \\ 1 & 1 & 1 & 0 & \ldots \\ 1 & 1 & 1 & 1 & \ldots \\ & & \vdots & & \end{bmatrix}$$

The response of the system compensated with only a fixed compensator to a unit step input is given by

$$C'(z) = K(z)R(z) = (3z^{-1}-2z^{-2})/(1-z^{-1})$$

from which, by long division, the following is obtained

$\mathbf{c}' = \text{col}\,[0, 3, 1, 1, 1, 1, \ldots]$

and observe that while the steady-state behavior is satisfactory, there is a 200 per cent overshoot in the second output sample. We now proceed to design a time-varying amplifier to reduce this overshoot.

Since equation 22 must be satisfied, the output cannot be specified at will. Let the desired output be denoted by

$\mathbf{c} = \text{col}\,[c_0, c_1, c_2, c_3 \ldots]$

Then

$\mathbf{c} - \mathbf{c}' = \text{col}\,[c_0, c_1 - 3, c_2 - 1, \ldots]$

and, from equation 19

$$\mathbf{q} = \begin{bmatrix} 1 & 0 & 0 & 0 & \ldots \\ 1 & 1 & 0 & 0 & \ldots \\ 1 & 1 & 1 & 0 & \ldots \\ 1 & 1 & 1 & 1 & \ldots \\ & & \vdots & & \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 - 3 \\ c_2 - 1 \\ c_3 - 1 \\ \vdots \end{bmatrix} = \begin{bmatrix} c_0 \\ c_0 + c_1 - 3 \\ c_0 + c_1 - 3 + c_2 - 1 \\ c_0 + c_1 - 3 + c_2 - 1 + c_3 - 1 \\ \vdots \end{bmatrix}$$

and

$\mathbf{e} = \mathbf{r} - \mathbf{c} = \text{col}\,[1 - c_0, 1 - c_1, 1 - c_2, \ldots]$

We thus have the following set of relations

$a(0) = c_0/(1 - c_0)$
$a(1) = (c_0 + c_1 - 3)/(1 - c_1)$
$a(2) = (c_0 + c_1 + c_2 - 4)/(1 - c_2)$
$a(3) = (c_0 + c_1 + c_2 + c_3 - 5)/(1 - c_3)$
$\vdots$

The simplest compensation is achieved by letting $c_0 + c_1 = 3$ and $c_2 = c_3 = c_4 = \ldots = 1$. In this case $a(1) = a(2) = \ldots = 0$ and $a(0) = c_0/(1 - c_0)$ can be set.

One of infinitely many possibilities is to set $c_0 = 1.5$. Then $c_1 = 1.5$ and $a(0) = -3$. Thus, by the addition of an amplifier whose gain is -3 at $t = 0$ and is then turned off, the overshoot has been reduced from 200 per cent to 50 per cent, as shown in Fig. 7.

Note that in this example it is possible to produce an output at the instant the input is applied, because the plant does not incorporate an integration in its transfer function. In a more realistic problem, the plant would contain an integration and it would be physically impossible to produce an output at the instant the input is applied. The foregoing analysis would reveal this by the presence of zero in the denominator of the expression for $a(0)$, which would require either the gain of the time-varying amplifier to be infinite or the output to be zero at the first sampling instant.

From the physical standpoint, the function of the amplifier is to provide suitably controlled supplementary signal to the plant during the transient period. The digital compensator provides steady-state correction.

Summary and Conclusions

The method described in the foregoing furnishes a simple routine method for the analysis of time-varying sampled-data systems. Since only two simple mathematical operations are involved, matrix multiplication and solution of single linear algebraic equations, they can be performed by relatively unskilled personnel or by digital computers. The disadvantages of the method are:

1. The transmission matrices for each component of the system must be known. For many systems this is not a serious shortcoming, since frequently the system may be regarded as comprising only fixed elements and amplifiers (or attenuators) with time-varying gain.

2. Steady-state behavior cannot be determined in general except by evaluating infinitely many elements of the output signal vector. In a few cases the trend toward steady-state behavior is seen after a few output samples have been computed, as in the first numerical example. In many systems, however, the question of steady-state behavior is of small concern, since the process involved is "terminal": after a finite and frequently small number of samples the system is turned off or reset to some initial state.

References

1. A Mathematical Technique for the Analysis of Linear Systems, J. R. Ragazzini. *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 42, Nov. 1954, pp. 1645–51.

2. Frequency Analysis of Variable Networks, L. A. Zadeh. *Ibid.*, vol. 38, Mar. 1950, pp. 293–99.

3. The Analysis of Sampled-Data Systems, J. R. Ragazzini, L. A. Zadeh. *AIEE Transactions*, vol. 71, pt. II, Apr. 1952, pp. 225–34.

4. A New Linear Operational Calculus, F. W. Bubb. *Technical Report No. 6581*, Wright Air Development Center, Dayton, Ohio, May 1951.

5. The Pulse Transfer Function and Its Application to Sampling Servo Systems, R. H. Barker. *Proceedings*, Institution of Electrical Engineers, London, England, pt. IV, Monograph no. 43, July 1952.

6. Analysis and Synthesis of Sampled-Data Control Systems, Eliahu I. Jury. *AIEE Transactions*, vol. 73, pt. I, Sept. 1954, pp. 332–46.

7. A Simplified Method of Solving Linear and Non-Linear Systems, R. Boxer, S. Thaler. *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 44, Jan. 1956, pp. 89–101.

8. Sampled-Data Processing Techniques for Feedback Control Systems, Arthur R. Bergen, John R. Ragazzini. *AIEE Transactions*, vol. 73, pt. II, Nov. 1954, pp. 236–47.

Discussion

B. H. Bharucha and **F. W. Semelka** (University of California, Berkeley, Calif.): This is an interesting approach to the analysis of linear time-varying sampled-data systems. The technique presented is based on the assumption that the output is required only at the sampling instants. However, in most applications of sampled-data systems the response between the sampling instants is of interest and should be investigated.

The purpose of this discussion is to extend the method of the paper to obtain the continuous response, thus removing the author's restriction and considerably increasing the value of the matrix method introduced. Furthermore, it is shown that synthesis based on system response only at the sampling instants may lead to misleading conclusions about the system's actual behavior. This is illustrated by examining more critically the author's synthesis example.

In the following derivations, it is assumed that the system is linear, time invariant between sampling instants and time-varying only at sampling instants.

Under the transformation

$$t = (n+m)T, \quad n = 0, 1, 2, \ldots \qquad 0 \le m \le 1 \tag{25}$$

the response $h(t,k)$ of a linear component to an impulse occurring at $t = kT$ may be written as

$$h(t,k) = h[(n+m)T, k] = h_n(m,k) \tag{26}$$

where $h_n(m,k)$ is defined as the piecewise time description of the impulse response $h(t,k)$.

Using the superposition theorem, and noting that for relaxed systems

$$h(m,k) = 0 \quad \text{for } k > n \tag{27}$$

the following obtains

$$y_n(m) = \sum_{k=0}^{n} h_n(m,k)x(k) \tag{28}$$

It is noticed that $y_n(m)$ describes the total response, in the $(n+1)$ sampling interval, of the component H to the input $\sum_{k=0}^{n} x(k)\,\delta(t - kT)$.

Equation 28 may be viewed as a matrix equation

$$\mathbf{y_n(m)} = \mathbf{H_n(m,k)x(k)} \tag{29}$$

where

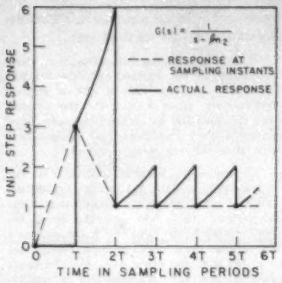


Fig. 8. Unit step response of system in Fig. 6(B). Output given by equation 41

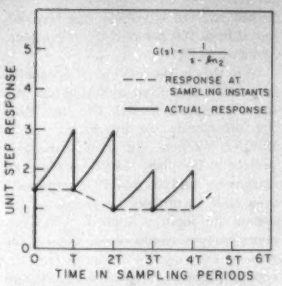


Fig. 9. Unit step response of system in Fig. 6(C). Plot of equation 43

$$\mathbf{y_n(m)} = \text{col}\,[y_0(m),\ y_1(m),\ \ldots,\ y_n(m)] \quad (30)$$

$$\mathbf{x(k)} = \text{col}\,[x(0),\ x(1),\ \ldots,\ x(n)] \quad (31)$$

and

$$\mathbf{H_n(m,k)} = \begin{bmatrix} h_0(m,0) & 0 & 0 & \ldots & 0 \\ h_1(m,0) & h_1(m,1) & 0 & \ldots & 0 \\ h_2(m,0) & h_2(m,1) & h_2(m,2) & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ h_n(m,0) & h_n(m,1) & h_n(m,2) & \ldots & h_n(m,n) \end{bmatrix} \quad (32)$$

It is of interest to note that for $m=0$, this matrix reduces to the transmission matrix for the sampled output given by the author.

If the component is fixed

$$h_n(m,k) = h_{n+b}(m,k+b), \quad \begin{matrix} b = \text{integer} \\ t = (n+m)T \end{matrix} \quad (33)$$

i.e., the characteristic response is independent of the time origin. In particular, for $k=0$, equation 33 gives

$$h_n(m,0) = h_{n+b}(m,b) \quad (34)$$

Let

$$h_n(m,0) = h_n(m) \quad (35)$$

Applying equations 34 and 35 to the matrix in equation 32 yields the transmission matrix $\mathbf{H_n(m)}$ of a linear fixed component.

$$\mathbf{H_n(m)} = \begin{bmatrix} h_0(m) & 0 & 0 & \ldots & 0 \\ h_1(m) & h_0(m) & 0 & \ldots & 0 \\ h_2(m) & h_1(m) & h_0(m) & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ h_n(m) & h_{n-1}(m) & h_{n-2}(m) & \ldots & h_0(m) \end{bmatrix} \quad (36)$$

Since the system is fixed and linear, $H(s)[=\mathcal{L}\{h(t)\}]$ is generally known, and the modified z-transform[1] of $H(s)$, $H^*(z,m)$, can be used to obtain the elements of $\mathbf{H_n(m)}$; $zH^*(z,m)$ can be expanded[2,3] in a Laurent series about $z=0$.

$$zH^*(z,m) = h_0(m) + h_1(m)z^{-1} + \ldots + h_j(m)z^{-j} + \ldots,\ |z|>1 \quad (37)$$

The coefficients of the series are then identified with the elements $h_n(m)$ of the transmission matrix $\mathbf{H_n(m)}$.

To illustrate the preceding method, the system of Fig. 6(B) solved by the author is re-examined. Since the continuous transfer function of the plant $G(s)$ was not specified by the author, one can choose several forms of $G(s)$ which have the specified z-transform $G^*(z)=1/1-2z^{-1}$. Selecting the simplest form of $G(s)$ yields

$$G(s) = \frac{1}{s-1n2}$$

The modified z-transform of $G(s)$[1] is

$$G^*(z,m) = \frac{2^m}{z-2}$$

$zG^*(z,m)$ may be expanded as

$$zG^*(z,m) = 2^m + 2^{m+1}z^{-1} + 2^{m+2}z^{-2} + \ldots + 2^{m+j}z^{-j} + \ldots \quad (38)$$

to yield

$$\mathbf{G_n(m)} = \begin{bmatrix} 2^m & 0 & 0 & \ldots & 0 \\ 2^{m+1} & 2^m & 0 & \ldots & 0 \\ 2^{m+2} & 2^{m+1} & 2^m & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 2^{m+j} & 2^{m+j-1} & \ldots & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & 2^m \end{bmatrix} \quad (39)$$

The output matrix $\mathbf{C_n(m)}$ is given by the following

$$\mathbf{C_n(m)} = \mathbf{G_n(m)D}\,[1+\mathbf{G_n(0)D}]^{-1}\mathbf{R} \quad (40)$$

For a unit step input, the solution of equation 40 is

$$\mathbf{C_n(m)} = \text{col}\,[0,\ 3(2^m),\ 2^m,\ 2^m,\ \ldots,\ 2^m,\ \ldots] \quad (41)$$

The output given by this matrix is shown in Fig. 8.

Similarly, applying the method to the time-varying system, Fig. 6(C), synthesized by the author, the continuous output is found to be

$$\mathbf{C_n(m)} = \mathbf{G_n(m)}\,[\mathbf{D}+\mathbf{A}]\,[1+\mathbf{G_n(0)}(\mathbf{D}+\mathbf{A})]^{-1}\mathbf{R} \quad (42)$$

For $G(s)=1/s-1n2$, the unit step response is given by

$$\mathbf{C_n(m)} = \text{col}\,[3(2^{m-1}),\ 3(2^{m-1}),\ 2^m,\ 2^m,\ \ldots,\ 2^m,\ \ldots] \quad (43)$$

a plot of which is shown in Fig. 9. The necessity of considering the total response is quite apparent.[4]

The shortcomings associated with methods considering the sampled values only can be further illustrated by the choice of another form of $G(s)$, corresponding to the given $G^*(s)$ for the plant in Fig. 6(C).

$$G(s) = \frac{s-1n2}{(s-1n2)^2+4\pi^2}, \quad G^*(s) = \frac{1}{1-2z^{-1}} \quad (44)$$

$$G^*(z,m) = \frac{2^m \cos 2\pi m}{z-2}$$

Solving for the output yields

$$\mathbf{C_n(m)} = \text{col}[3(2^{m-1})\cos 2\pi m,\ 3(2^{m-1})\times\cos 2\pi m,\ 2^m\cos 2\pi m,\ 2^m\cos 2\pi m,\ \ldots,\ 2^m\cos 2\pi m,\ \ldots] \quad (45)$$

A comparison of the true output and the sampled output, Fig. 10, demonstrates that the designer gains only a superficial knowledge of system behavior[2] by studying the sampled response exclusively.

The purposeful introduction of a pure time-varying compensator (controlled solely by a clock source) is useful to improve system response only in cases where the input is completely specified for all time values of interest. But such systems are not frequently encountered. Consequently, we believe that when time-varying compensation is used and the exact input is unknown, the only meaningful design would be one which based the time-varying compensation further on the various signals present in the system.

In conclusion, we believe that the use of the matrix representation is a convenient and useful method for analysis of sampled-data systems (both fixed and time-varying). The matrix method may also prove on further investigation to be of value in synthesis.

REFERENCES

1. Synthesis and Critical Study of Sampled-Data Control Systems, Eliahu I. Jury. *AIEE Transactions*, vol. 75, pt. II, July 1956, pp. 141–51.

2. Hidden Oscillations in Sampled-Data Control System. *Ibid.* (Jan. 1957 Section), pp. 391–95.

3. A Method of Determining the Coefficients of the Modified z-Transform Expansion, E. I. Jury, F. W. Semelka. *Internal Memorandum No. 1*, Sampled-Data Systems Group, Electronics Research Laboratories, University of California, Berkeley, Calif., Aug. 1956.

4. *Class notes*, Electrical Engineering 208–5, Sampled-Data Control System, University of California, Spring 1956.

Jack Sklansky (RCA Laboratories, Princeton, N. J.): Although the paper is titularly an analysis, it does offer a form of synthesis technique in its section entitled, "Improvement of Feedback System Performance by Means of Time-Varying Amplifier." The following question, neglected in the paper therefore arises: How does one justify the intentional use of a time-varying component in a control system?

This question becomes particularly urgent when one notes that as far as its input-output relations are concerned, the system synthesized in the second numerical example is not truly time-varying, even though it contains a time-varying amplifier. The reason for this is that the time variation of the variable amplifier is tacitly assumed in the paper to be synchronized with the time of application of the input step function, so that a delay nT in the application of the input will merely delay the output by nT without changing its size or shape.

It seems to me that the general reasons for the intentional use of time-varying components are:

1. To demodulate or otherwise decode the input, and/or to modulate or otherwise encode the output so that they may be used or transmitted more efficiently. (The sampling process itself usually has this purpose.)

2. To compensate for an unavoidable time-variation in one of the components or in the environment.

3. To produce a system whose overall behavior is time-invariant, but which, because of economic or behavioral factors, is better than equivalent systems constructed with linear components throughout.

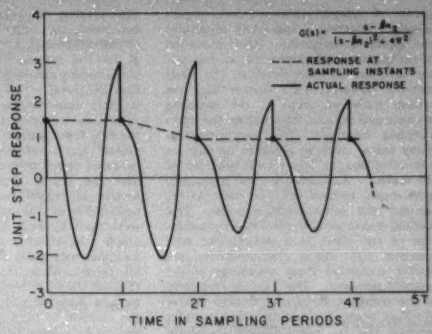
If one were to synthesize a system having the same over-all transfer function and the same $G(s)$ as the final system of the second numerical example, but having no time-varying component, then the required $D^*(s)$ would have a zero that cancels the unstable pole of $G^*(z)$. Because in practice such cancellation cannot occur exactly, the resulting system will be unstable.[1] Thus, the justification of the system of the second numerical example is given by the third of the reasons listed in the foregoing.

A practical difficulty seems to occur in the input-synchronization of the time-varying amplifier of Fig. 6 when the input is a ramp or any other time function having an initial value of zero, and when only the sampled values of the input are available for synchronization. In such a case the amplifier will be turned on one or more sampling periods later than it would be if the input had a very small pulse superimposed on it at $t=0$.

The following techniques are suggested for determining the steady-state response of linear time-varying sampled-data systems. Only inputs of the form $e^{j\omega t}$ will be considered, since cases involving other forms of periodic inputs may be handled by expanding the inputs into their complex Fourier series.

First, suppose the time variation of the variable components is "transient," as in the case of the author's second numerical example; i.e., suppose the time variation eventually stops or dies out. Then the effect of time variation of the variable component can be obtained by superimposing an equivalent transient signal on the input and replacing the variable component by the proper time-invariant device. The component of response corresponding to this transient signal will die out and, since the system is linear, the transient signal will have no effect on the steady-state response. Hence, when the time variation is transient, the steady-state output of the actual system is equal to the steady-state output of that time-variant system obtained by replacing the variable components by the time-invariant components they become or approach, as $t\to\infty$. Thus, if the system transfer function approaches $K^*(z)$ as $t\to\infty$, then the steady-state response of the system to $e^{j\omega t}$ is just $K^*(e^{j\omega T})e^{j\omega nT}$.

Fig. 10. Unit step response of system in Fig. 6(C). Comparison of true and sampled outputs



Second, suppose the time variation of the variable components is periodic. Then, when the input is $e^{j\omega t}$, the system's response at the sampling instants is

$$c(nT)=e^{j\omega nT}\sum_{m=0}^{n}k_{n-m,m}e^{-j\omega mT} \tag{46}$$

where k_{ij} is the output of the system at $t=(i+j)T$ in response to a single pulse of unit height, covering a single sampling instant, applied to the input terminals at $t=iT$. Let $K_n^*(z)$ be defined as that periodic function of n approached by

$$\sum_{m=0}^{n}k_{n-m,m}z^{-m} \tag{47}$$

as n approaches infinity, considering n for the moment as a continuous variable rather than discrete. Using this definition, and taking note of equation 46, it is seen that the steady-state component of $c(nT)$ in response to $e^{j\omega t}$ is formally given by

$$c_{ss}(nT)=e^{jn\omega T}K_n^*(e^{j\omega T}) \tag{48}$$

For example, if

$$K_n^*(z)=\frac{a\cos n\omega_1 T}{1-(b\sin 2\,n\omega_1 T)z^{-1}} \tag{49}$$

then, in response to $e^{j\omega t}$, the steady-state output is

$$c_{ss}(nT)=\left[\frac{a\cos n\omega_1 T}{1-(b\sin 2n\omega_1 T)\,e^{-j\omega T}}\right]e^{jn\omega T} \tag{50}$$

Unfortunately, a closed-form expression is not always possible. Such a situation may occur in a feedback system where $K_n^*(z)$ must be expressed in terms of the inverse of (1+loop transfer function). In this case $K_n^*(z)$ may be computed by using the method given in the paper of inverting triangular matrices, and making the approximation

$$K_n^*(z)\cong\sum_{m=0}^{M}k_{n-m,m}z^{-m},\quad M<n \tag{51}$$

where M is taken large enough to make the error acceptably small.

REFERENCE

1. FACTORS IN THE DESIGN OF DIGITAL CONTROLLERS FOR SAMPLED-DATA FEEDBACK SYSTEMS, John E. Bertram. *AIEE Transactions*, vol. 75, pt. II, July 1956, pp. 151–59.

John M. Salzer (The Magnavox Company, New York, N. Y.): The transient solution of actually or artificially sampled systems is particularly amenable to systematization. The author has carried this systematization further by applying matrix algebra to the analysis and design of time-variant systems.

The use of matrix algebra adds considerable facility and clarity to dealing with this problem. Previous work (see reference 7 of the paper) shows how to solve time-variant differential equations. In this case the starting point is a relation between the controlled and reference quantities. The use of the matrix method conveniently allows the time variation to be specified for a particular element of the system from which the closed-loop transfer relationship can be systematically obtained. The clarity provided by the use of matrix methods further permits the design of specific compensations to attain desired system performance. The choice of a time-variant amplifier for improving a particular system is a noteworthy example of using the new method. Further exploitation of the method is expected to add to the increasing store of new methodology in system studies.

Bernard Friedland: Mr. Bharucha and Mr. Semelka are quite correct in their assertion that the problem of determining the re-

sponse of sampled-data systems between sampling intervals is frequently important and merits investigation. Their extention of the matrix method provides an interesting solution to this problem. In their method, however, they assume that the system varies in steps at the sampling instants and remains fixed during the intersampling intervals. This assumption, in many cases, must be regarded as only an approximation to the actual behavior of the system. An alternative and frequently more rapid method may be used to determine the intersample response. This method makes use of the fact that the excitation of the plant is a staircaselike time function $f(t)$ whose discontinuities are readily computed from knowledge of the sampled response. The Laplace transform of the continuous response is thus $C(s) = G(s)F(s)$ where $G(s)$ is the plant transfer function and $F(s)$ is the Laplace transform of $f(t)$.

I also agree that a meaningful design must consider all possible inputs to the system. Design techniques which consider the ensemble of signals which comprises the input to the system are essentially statistical and are outside the scope of the paper. However, since the transmission matrix of the compensated system is known, it is possible to compute the response of the system (at sampling intervals) to any specified input, and thereby assess its performance.

Mr. Sklanky's remark that the system discussed in the second example is not truly time-varying is puzzling. As originally conceived, the time scale is independent of the input. If the transmission matrix **K** for the system compensated with the time-varying amplifier is written, it is found that $k(i,j) \neq k(i-j)$ and hence the system is truly time-varying. It is possible, of course, to synchronize the time scale with the first input sample. However, once the "clock" starts running it can be reset only by some means independent of the system and the input. Otherwise the system matrix **K** would not give a complete description of the system. If the clock were reset by each input sample, for example, the effect would be to have an amplifier with fixed gain in the system at all times. While this system would be time-invariant, it would not behave in the manner intended. If, on the other hand, the decision to reset the clock is made to depend upon some signal in the system, the behavior is nonlinear.

The paper did not consider steady-state response because it was felt that, in all but the most obvious cases, the use of matrices, which become infinite as the steady state is approached, has practical limitations, and that extending the z-transform techniques used for fixed systems would probably be more convenient. Investigation conducted since the paper was written has substantiated this. In particular, one result of the investigation may be expressed as follows: Let $K(n,z)$ be the z-transform of the nth row of the transmission matrix, and let

$$K(\infty,z) = \lim_{n \to \infty} K(n,z)$$

if this limit exists. If the poles of $K(\infty,z)$ lie within the unit circle of the z-plane, then: 1. the response to a bounded input is bounded, and 2. the steady-state response is given by

$$c(\infty) = \lim_{z \to 1} (1 - z^{-1}) K(\infty,z) R(z)$$

where $R(z)$ is the z-transform of the input. These results constitute an extension of Mr. Sklanky's method of determining steady-state response. It is also interesting to note that the stability criterion, given by item 1 in the foregoing, can be obtained from the theory of infinite matrices.[1] The difficult question in connection with time-varying sampled-data systems which has not been completely answered is whether $K(\infty,z)$ exists and, what its value is, given the system configuration and the transmission functions $H(n,z)$ of its components. It is clear that if all time variation ceases after a finite number of samples, then $K(\infty,z)$ is simply the pulse transfer function of the resulting fixed system. However, if one or more of the components in the system approach, but never actually achieve, time invariance, then the form of $K(\infty,z)$ is not altogether clear.

REFERENCE

1. INFINITE MATRICES AND SEQUENCE SPACES (book), R. G. Cooke, Macmillan & Company, Ltd., London, England, 1950.

Variably Damped Servomechanisms

L. SCHIEBER
ASSOCIATE MEMBER AIEE

THE SPEED of response is one of the most important characteristics of a large class of servomechanisms. A variety of methods and criteria for determining the speed of response exists. One of these criteria is the length of time between the application of a step of input to the servo and the time when the output corresponds to the new input, within a specified tolerance. This length of time is usually referred to as the settling time. Fig. 1 shows the settling time of a typical underdamped servo. In most applications, the designer desires to obtain as small a settling time as possible, i.e., he will design the servo with the highest possible speed of response.

The responses to step inputs of three typical servos are shown in Fig. 2. Curve *A* shows the response of a lightly damped servomechanism. While this servo reaches the new value of output quickly, it oscillates outside of the tolerance limits for some time. This results in a rather large settling time. Curve *B* shows the response of a highly damped servomechanism. While the oscillations of servo *A* are damped out, the speed of response has been reduced so much that the settling time is again very large. Curve *C* represents a compromise between these two extremes. Servo *C* approaches the new value of output more rapidly than servo *B* and less rapidly than servo *A*. However, servo *C* settles within the specified tolerance more quickly than either of the other two. This is the compromise that most servo designers seek.

The response of servo *C* represents a compromise between the desired characteristics of a large initial slope and a rapid decay of the oscillatory part of the response. Curve *D* in Fig. 3 shows that a servo following the response curve of servo *A* to point *c* and then switching to curve *B* would have a smaller settling than servo *C*. This approach has been applied to contactor servos by McDonald.[1] The response shown by curve *E* in Fig. 3 uses the initial slope of servo *A*, the intermediate slope of servo *C*, and the slope of servo *B* to improve the response even further. By taking smaller sections of a greater number of curves, an optimum response having a minimum settling time can be approached. This is done in the servo described in this paper by making the damping of the servo continuously variable and an inverse function of the magnitude of the error.

Another approach to the measurement of the speed of response of a servomechanism is its response to a ramp input. The response of a typical servo to a ramp input is shown in Fig. 4. The steady-state error shown in Fig. 4 is

$$E_{ss} = \frac{\omega_i}{K_v} \qquad (1)$$

This relationship defines the velocity

Paper 56-958, recommended by the AIEE Feedback Control Systems Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Fall General Meeting, Chicago, Ill., October 1–5, 1956. Manuscript submitted October 21, 1954; made available for printing July 23, 1956.

L. SCHIEBER is with the Kearfott Company, Clifton, N. J.

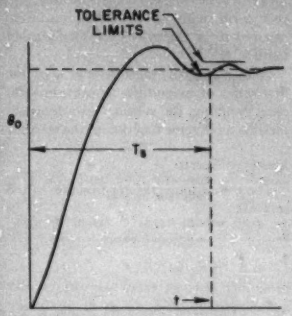


Fig. 1. Settling time of a typical underdamped servo

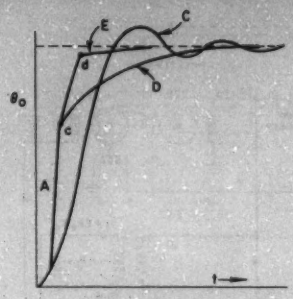


Fig. 3. Improvement of the response through switching

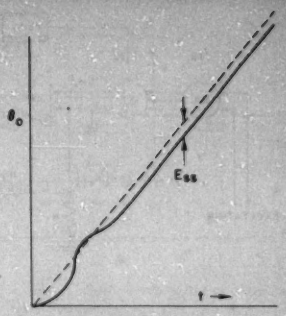


Fig. 4. Response of an underdamped servo to a ramp input

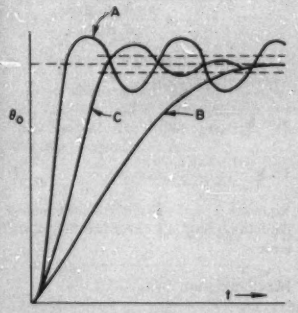


Fig. 2. Settling times of three types of servos

error constant K_v of the servo. The settling time is a function of K_v. However, the velocity error constant as defined by equation 1 has a more direct physical meaning. A large velocity error constant will enable a servo to follow high velocity inputs with a small error. The servo described in this paper has a velocity figure of merit (no longer a constant) which is roughly proportional to the input velocity. Thus, the velocity error constant is larger for higher velocity inputs. This ensures a smaller steady-state error for high velocity inputs.

The desirability of high-performance servomechanisms has been discussed as if high performance were easily attained. However, to be more practical, even when servos are designed with suitable compensation to make them appear satisfactory in an analysis, building this high performance in the servo is complicated by many considerations. These include saturation of the amplifier by noise or phase shift in the transducers, a limitation of the gain of the amplifier caused by inherent noise, and a reduction in stability of the servo by backlash in the gearing. From a practical standpoint, the problem then resolves itself into one of designing a high-performance servomechanism with a low amplifier grain which can withstand a reasonable amount of backlash in the gearing, and amplifier gain variation without affecting its stability. The use of variable damping offers a solution to this problem.

Nomenclature

E = error, rad (radians)
E_{ss} = steady-state error, rad
g = gear ratio
K_a = amplifier gain
$K_D = K_1/R$—proportionality constant
K_s = error transducer gain, volts/rad (volts per rad)
K_m = motor gain, rad/volt sec (second)
K_T = tachometer sensitivity, sec/rad
$K_T' = K_T K_1$, volt sec/rad
K_v = velocity error figure of merit, sec^{-1}
K_1 = proportionality constant
s = Laplace operator, sec^{-1}
T_m = motor time constant, sec
T_s = settling time, sec
V_1 = fixed voltage, volts
V_m = nominal control field voltage, volts
V_r = hypotenuse voltage, volts
θ_i = input shaft position, rad
θ_o = output shaft position, rad
ζ = damping ratio
ω_i = input velocity, rad/sec
ω_n = undamped natural frequency, rad/sec
θ_m = motor shaft position, rad
G_m = frequency dependent portion of motor transfer function, sec

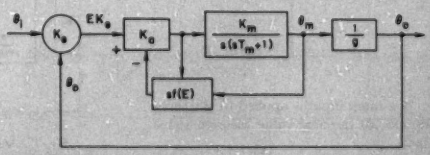


Fig. 5. Block diagram of a variably damped servomechanism

Philosophy

A variably damped servo is one in which the damping of the servo is inversely proportional to the magnitude of the error existing at any time. In this manner, the characteristics of the servo are varied to meet the needs imposed by the input. Thus, when a step function is applied and a large error suddenly develops, the damping of the servo approaches zero, allowing the motor to approach its rated speed. As the error decreases, signifying an approach to correspondence of the output with the input, the damping increases sufficiently to ensure stable operation at the new position. The maximum speed of the motor may be utilized during most of the response provided the damping increases rapidly near the null position. The settling time may be made considerably less than that of a conventional servo.

If a ramp input is applied to the servo, an analogous transient response will take place. However, some steady-state error will exist after the transient dies out. In a conventional servo, the magnitude of the steady-state error would be proportional to the input velocity. However, the damping is inversely proportional to the error in a variably damped servomechanism. Therefore the velocity error constant will be directly proportional to the error. The increased velocity error constant will mean that the variably

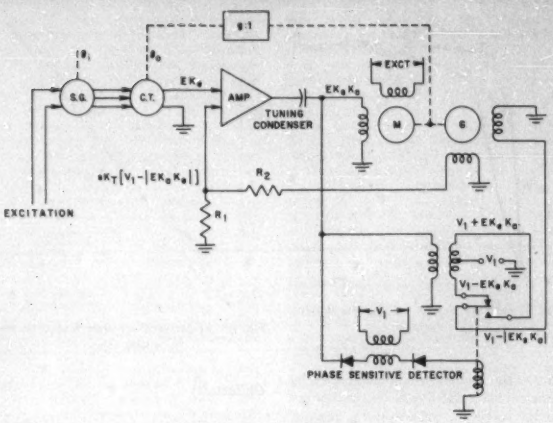


Fig. 6. Variably damped position servo

damped servo will have less following error than the conventional servo.

Derivation of Equations

The use of variable damping in a second-order system is illustrated in the block diagram of Fig. 5. Compensation is obtained through the use of tachometric feedback. The variable damping feature is introduced through the fixed field of the tachometer. The details of the circuit will be discussed later.

Velocity Error Figure of Merit

The open-loop transfer function of the servo shown in Fig. 5 is

$$\frac{\theta_0}{E}=\frac{K_cK_aK_m}{g[1+K_aK_mf(E)]}\times\left[\frac{1}{s\left(s\frac{T_m}{1+K_aK_mf(E)}+1\right)}\right] \quad (2)$$

Substituting (θ_i-E) for θ_0 and solving for E gives

$$E=\frac{\theta_i}{1+\frac{K_cK_aK_m}{sg[1+K_aK_mf(E)]}\times\left[\frac{1}{s\frac{T_m}{1+K_aK_mf(E)}+1}\right]} \quad (3)$$

If a ramp input is applied to the servo $(\theta_i=\omega_t/s^2)$ the final value theorem states that

$$E_{ss}=\lim_{s\to 0}\frac{\omega_t}{s+\frac{K_cK_aK_m}{g[1+K_aK_mf(E)]}\times\left[\frac{1}{s\frac{T_m}{1+K_aK_mf(E)}+1}\right]}=\frac{\omega_t}{\frac{K_cK_aK_m}{g[1+K_aK_mf(E)]}} \quad (4)$$

Referring to the usual definition of the velocity error constant, equation 1

$$K_v=\frac{K_cK_aK_m}{g[1+K_aK_mf(E)]} \quad (5)$$

The term velocity error constant is a misnomer in the present case; this expression is obviously not a constant since it is a function of the error. Therefore this term will be more appropriately referred to as the velocity error figure of merit.

Since $K_aK_mf(E)$ must be of at least the first order of magnitude to provide sufficient damping, the velocity error figure of merit is an inverse function of the error.

Damping Ratio

The closed-loop transfer function of the servo is

$$\frac{\theta_0}{\theta_i}=\frac{K_cK_aK_m}{sg\left\{[1+K_aK_mf(E)]\times\left[s\frac{T_m}{1+K_aK_mf(E)}+1\right]\right\}+K_cK_aK_m} \quad (6)$$

$$=\frac{1}{s^2\frac{gT_m}{K_cK_aK_m}+sg\frac{[1+K_aK_mf(E)]}{K_cK_aK_m}+1}$$

Using the standard definitions for a second order system[2]

$$\omega_n=\sqrt{\frac{K_cK_aK_m}{gT_m}} \quad \text{and} \quad (7)$$

$$\zeta=\sqrt{\frac{g[1+K_aK_mf(E)]^2}{4K_cK_aK_mT_m}} \quad (8)$$

Equation 8 shows that the dimensionless damping ratio ζ is a direct function of the error.

Mechanization

The dependency of the tachometer feedback on the magnitude of the error may be obtained in a number of ways. Two methods of obtaining $f(E)$ will be discussed in this section.

The first method is shown schematically in Fig. 6. For this mechanization, $f(E)=K_T(V_1-|EK_aK_b|)$. Since E may be either positive or negative, a phase-sensitive detector and relay are used to obtain the absolute magnitude of EK_aK_b.

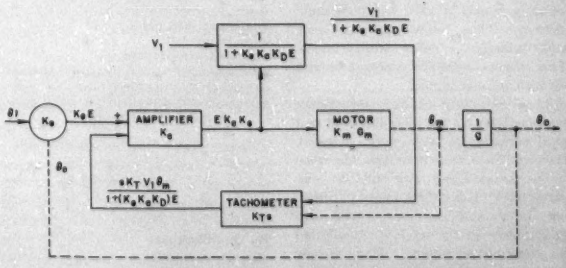


Fig. 7. Variably damped servomechanisms with damping control amplifier

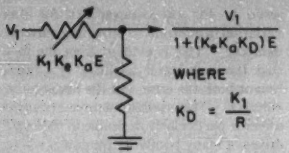


Fig. 8. Functional schematic of damping control amplifier

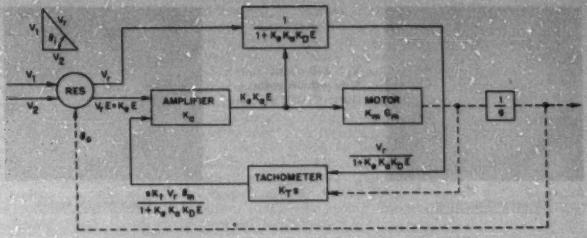


Fig. 10 (right). Tangent-solving servo with variable damping

(If a polarized relay is used, the phase-sensitive detector may be eliminated.) Substituting this $f(E)$ in equations 5 and 8 obtains the following results

$$K_v = \frac{K_s K_a K_m}{g[1+K_a K_m K_T(V_1 - |EK_a K_s|)]} \tag{9}$$

$$\zeta = \sqrt{\frac{g[1+K_a K_m K_T(V_1 - |EK_a K_s|)]^3}{4K_s K_a k_m T_m}} \tag{10}$$

The relationship $EK_a K_m K_s = \theta_m = g\theta_0 = g\omega_t$ may be used to write K_v and ζ as functions of the input velocity. The resulting expressions are

$$K_v = \frac{K_s K_a K_m}{g\left[1+K_a K_m K_T'\left(1-\frac{g\omega_t}{K_m V_1}\right)\right]}$$

where $K_T' = K_T V_1$ (11)

$$\zeta = \sqrt{\frac{g\left[1+K_a K_m K_T'\left(1-\frac{g\omega_t}{K_m V_1}\right)\right]^3}{4K_s K_a K_m T_m}} \tag{12}$$

Denoting the voltage on the control field of the motor by V_m gives

$$\frac{g\omega_t}{K_m V_1} = \frac{V_m}{V_1}$$

When $V_m/V_1 = 1$, equations 11 and 12 reduce to

$$K_v = \frac{K_s K_a K_m}{g} \tag{11A}$$

$$\zeta = \sqrt{\frac{g}{4K_s K_a K_m T_m}} \tag{12A}$$

These are the maximum K_v and minimum ζ obtainable without allowing the feedback to become positive. Of course, there is no reason that the feedback should not be positive, provided the servo has enough damping to prevent oscillations at the maximum input velocity. In the case of a step input, damping should be sufficient to prevent oscillations at the slewing speed of the motor.

Fig. 7 shows a schematic of a similar servo where

$$f(E) = \frac{V_1 K_T}{1+K_s K_a K_D E}$$

This function is obtained from a damping control amplifier. The action of the amplifier is shown in the functional schematic of Fig. 8 and the circuit schematic is shown in Fig. 9. The control is obtained by varying the transconductance of the pentode as a function of the magnitude of the error. The expressions for K_v and ζ are, substituting in equations 5 and 8

$$K_v = \frac{K_s K_a K_m}{g\left[1+\frac{K_a K_m K_T V_1}{1+K_s K_a K_D E}\right]} \tag{13}$$

$$\zeta = \sqrt{\frac{g\left[1+\frac{K_a K_m K_T V_1}{1+K_s K_a K_D E}\right]^3}{4K_s K_a K_m T_m}} \tag{14}$$

Equations 13 and 14 can be written in

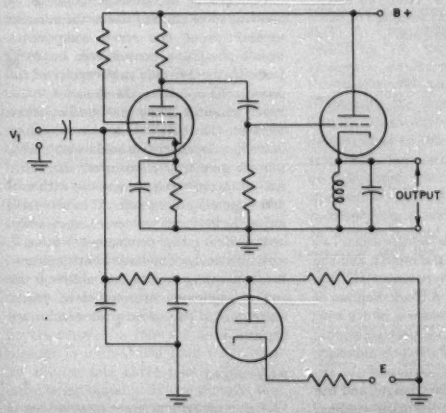


Fig. 9 (above). Circuit diagram of a damping control amplifier

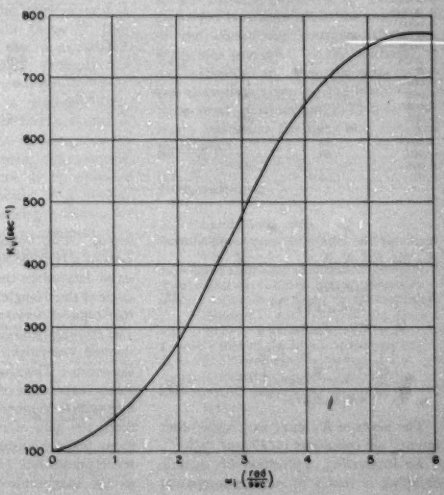


Fig. 11 (right). Velocity error figure of merit versus input velocity for a variably damped servo

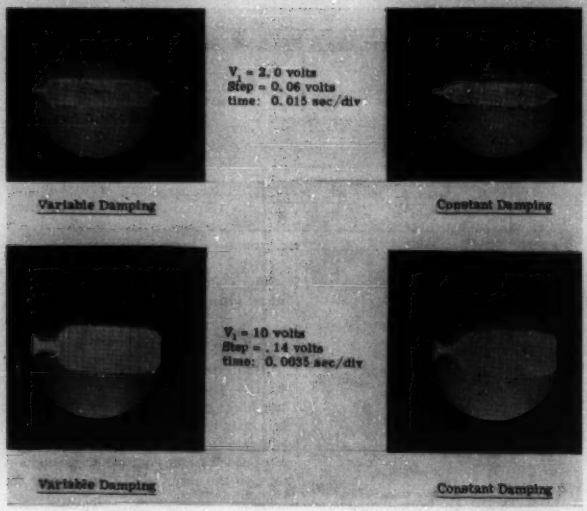


Fig. 12. Effect of variable damping on the transient response of a triangle solver

Table I. Velocity Error Figures of Merit of Variably Damped Tangent Solver

Velocity Error Figures of Merit, Sec^{-1}

Error, Radians	Hypotenuse Voltage, Volts						
	0.2	0.5	1.0	5.0	10.0	20.0	30.0
0.0	7.5	17.9	33.4	109	151	189	206
0.00021							439
0.00025						426	564
0.00033							884
0.0004				168			
0.0005						620	
0.0006					288		
0.0008					320		
0.001				176	331		
0.0015						737	
0.002		89	106	236			
0.0025							825
0.003		108			348		
0.004	32.5	120					
0.0045	58						
0.005	84		130				
0.006		148					
0.01	90		150				

terms of the input velocity, substituting $g\omega_i$ for $EK_aK_mK_s$:

$$K_s = \frac{K_sK_aK_m}{g\left[1+\frac{K_aK_m{}^2K_TV_1}{K_m+K_Dg\omega_i}\right]} \quad (15)$$

$$\zeta = \sqrt{\frac{g\left[1+\frac{K_aK_m{}^2K_TV_1}{K_m+K_Dg\omega_i}\right]^2}{4K_sK_aK_mT_m}} \quad (16)$$

The limits of K_s and ζ as E approaches infinity are equations 11(A) and 12(A).

An interesting application of variable damping is found in certain computing servos. Two voltages representing two sides of a triangle are the inputs, and the servo determines the hypotenuse and one angle of the triangle. A block diagram of this type of servo is shown in Fig. 10. The two inputs excite an electromagnetic resolver, one output of which is nulled by the servo. The shaft position of the resolver is the angle of the triangle, and the hypotenuse appears on the other winding. Since the gain of resolver (K_s) depends upon the magnitude of the inputs, the servo parameters (K_s) and (ζ) will vary as the magnitude of the inputs. The servo of Fig. 10 compensates for this variation in addition to providing variable damping. This is evident from the fact that the tachometer feedback is a function of both the error and the hypotenuse voltages. This variably damped triangle solver is particularly desirable in the optic drives of many bombing computers where high tracking speeds are required.

Experimental Results

Both of the servos previously described have been breadboarded and tested. A graph of velocity error figure of merit versus input velocity is given in Fig. 11 for the servo of Fig. 6. From the graph, the velocity error figure of merit is seen to increase more than seven times when the input velocity is increased to 5 rad/sec.

Table I shows velocity error figures of merit of a tangent solver as a function of both hypotenuse voltage and error. For example, assume a hypotenuse voltage of 0.2 volt and an error of 0.01 rad; the velocity error figure of merit would be 90. This is twelve times the velocity error figure of merit of the equivalent servo without variable damping. The transient response of this servo is compared to that of a conventional servo having the parameters that the variably damped servo has at zero error in Fig. 12. Note that the settling time has been considerably reduced.

Conclusions

In summing up, variable damping, by allowing more efficient use of the inherent capabilities of the servo components, makes possible a considerable improvement in the dynamic performance of the servo. This enables the designer to use lower amplifier gains and smaller servo motors. The stability of the servo can be made less dependent on backlash by using a lower gain error transducer; the variation of the servo characteristics with amplifier gain is quite small. The servo will follow large input velocities accurately, and an ideal transient response is possible with no overshoot and a low settling time. It is interesting to note that although the system equations are nonlinear, many techniques of the operational calculus are applicable.

References

1. Nonlinear Techniques for Improving Servo Performance, Donald McDonald. *Proceedings*, National Electronics Conference, Chicago, Ill., 1950.

2. Principles of Servomechanisms (book), G. S. Brown, D. P. Campbell. John Wiley & Sons, Inc., New York, N. Y. 1948.

Discussion

James R. Eckel, Jr. (University of Tennessee, Knoxville, Tenn.): The use of variable damping in a second-order positional servo using tachometric feedback was treated by J. B. Lewis in his paper, "The Use of Nonlinear Feedback to Improve the Transient Response of a Servomechanism," which appeared in *AIEE Transactions*, vol. 71, pt. II, Jan. 1953, pp. 449–53. It is surprising that no mention was made of this similar earlier investigation.

Gordon R. Slemon (University of Toronto, Toronto, Ont., Canada): The author has shown that variable damping can be used to achieve an improved step and ramp response in a servomechanism. To fulfill its function such a system must prevent an overshoot or must make the overshoot quite small. If a large overshoot does occur the system returns to its underdamped condition. This property of variably damped servomechanisms may be emphasized by determining the response of such a system to a pulse of sufficient magnitude to produce initial underdamping and of sufficient width to allow the system to achieve a considerably velocity. For this signal which might be a disturbance of some sort the system would be underdamped during most of its first major excursion and its maximum response to the disturbance might be very large.

In the application for which the author designed this variably damped system, it is true that its response will be better than can be achieved with a linear system. There are, however, other classes of input signals, especially those containing noise, for which its performance might prove to be greatly inferior to that of a linear system.

Response of a Particular Nonlinear Control System to Random Signals

P. N. NIKIFORUK
NONMEMBER AIEE

Synopsis: The response of an electromechanical control system, which incorporates by design a hard-spring amplitude characteristic, is evaluated for inputs having a Gaussian amplitude-probability distribution and no constant components. These inputs are considered not as noise, but as true signals. The analysis is limited to a system with a second-order characteristic equation and is based on a general extension of the describing function technique. The control system considered is simulated on an analogue computer and experimental verification of the analysis is obtained.

FOR a number of years, nonlinear control systems have been subjected to intensive experimental and theoretical studies. In the majority of cases these studies have been limited to the analysis of systems with inherent nonlinear characteristics. More recently interest has arisen in systems using nonlinear characteristics for producing distinctive and desirable features of system behavior.[1–3] A particular nonlinear system which was considered of interest is one incorporating the hard spring,[4] a frequency-invariant amplitude-dependent characteristic. Interest in this characteristic was aroused by the suggestion that the tensioning of muscles in animal and man may follow an approximate cubic relationship with stimulating signal.[5] It was thought that a system incorporating a similar characteristic might have some features.

Theoretical and experimental work,[6,7] on a system incorporating a hard-spring characteristic has revealed the presence of several distinctive features of system behavior. The first is that the introduction of this characteristic actually decreases the response time to step inputs. The second is that for a sinusoidal excitation a sharp discontinuity in the gain-frequency characteristic arises under certain conditions of signal amplitude and degree of damping. This discontinuity in the gain-frequency characteristic perseveres in the presence of noise, providing that the latter is not too great.[7,8] This suggests that by preamplifying the input by a controlled amount a similar system, properly designed, might be used as a very-low-pass filter when the input is relatively sinusoidal. To evaluate the merits of such a system, it is of interest to determine its response to random signals. The purpose of this paper, therefore, is to evaluate theoretically the response of a system incorporating a hard-spring nonlinearity to random signals and to compare this response with that obtained on a servo simulator.

Paper 56-924, recommended by the AIEE Feedback Control Systems Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Fall General Meeting, Chicago, Ill., October 1–5, 1956. Manuscript submitted March 26, 1956; made available for printing July 19, 1956.

P. N. NIKIFORUK is with the Canadian Armament Research and Development Establishment, Quebec, Que., Canada.

The author wishes to express his sincere appreciation to Dr. J. C. West of Manchester University under whose guidance much of this work was conducted.

The System

The system to be considered is shown in Fig. 1. It is of second order, is stabilized by phase advance, and incorporates a simple hard-spring amplitude nonlinearity of the type shown in Fig. 2. It is one of a large class in which the output torque is controlled by the processed error signal. The response of it, and of all other servomechanisms of this class, is characterized by the fact that harmonics and other distortion components generated by the nonlinear element are severely attenuated by the filtering action of the servomotor. In the theoretical analysis of such systems this inherent characteristic permits the nonlinearity to be approximated by a quasilinear amplifier, the gain of this amplifier being determined by the value of the signal at its input. Thus, for a sinusoidal excitation it is assumed in the evaluation of the closed-loop response that the harmonics generated by the nonlinearity are negligible, and that the response of this element is given by the first Fourier component of its output. This is commonly known as the describing function technique.[9]

Nomenclature

V_c = control signal level
Δ, g, T, T_d = constants of the system (see Fig. 1)
k_{eq} = equivalent gain of nonlinear element
$P_1(x)$ = first-probability density function
σ = rms value of the input to the nonlinear element
$\sigma_n = \sqrt{2}\sigma/V_c$ = normalized value of σ
I = rms value of the input to the system
$I_n = \sqrt{2}I/V_c$ = normalized value of I
Φ_0 = zero-frequency spectral density
Φ_{11} = spectral density of the input to the system
$\Phi_n = 2g^4\Phi/V_c$
$\omega_0 = \sqrt{g}/T$
$\xi = T_d\omega_0$
$\omega_n = \omega_0 T_f$
ω = frequency

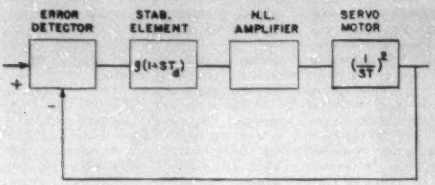
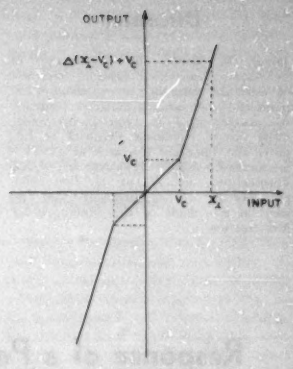


Fig. 1 (left). The basic system

Fig. 2 (right). The nonlinear characteristic



Extension of Describing Function Technique

There are two ways[10] of representing the response of a single-valued nonlinearity. The first is the single-function representation. For it the response is defined by

$$x_o = f(x_i) \tag{1}$$

where x_i is the instantaneous input to the nonlinear element and x_o is the instantaneous output of the nonlinear element.

The second is the 2-function representation and is defined by the following equation

$$x_o = k_{eq}x_i + x_H \tag{2}$$

where k_{eq} is the amplitude-dependent gain and x_H is the distortion component.

If the component x_H is to be neglected in the closed-loop analysis, the equivalent gain k_{eq} must be so chosen that the approximation is optimum in some respect. Booton[10] has shown that the describing function technique may be interpreted as a mean-square linearization and that k_{eq} should be that value which minimizes the mean-square error in the approximation. That this assumption is the same as that made for the describing function follows from the fact that the amplitude of each term in the Fourier series expansion of a function is such as to minimize the mean-square difference between that component and the function. Thus, for a sinusoidal input, k_{eq} is the quasilinear gain represented by the describing function.

For a particular value of the input, the squared error is

$$[f(x_i) - k_{eq}x_i]^2$$

The mean-square error E_m is obtained by averaging the squared error with the input to the element. The result is the equation

$$E_m = \int_{-\infty}^{\infty} [f(x_i) - k_{eq}x_i]^2 \, P_1(x_i)\,dx_i \tag{3}$$

where $P_1(x_i)$ is the first probability density function of the input.

The value of k_{eq} which minimizes E_m is obtained by expanding equation 3 and differentiating with respect to k_{eq}. The result is

$$k_{eq} = \frac{\int_{-\infty}^{\infty} x_i f(x_i) P_1(x_i)\,dx_i}{\int_{-\infty}^{\infty} x_i^2 P_1(x_i)\,dx_i} \tag{4}$$

Response of the Nonlinear Element

The signal at the input to the element is stationary. It has a Gaussian amplitude-probability distribution and no constant component. The first probability density function is therefore

$$P_1(x_i) = \frac{1}{\sigma\sqrt{2\pi}}\, \epsilon^{-\frac{x_i^2}{2\sigma^2}} \tag{5}$$

where σ is the rms value of the input to the nonlinear element.

The characteristic equation of the nonlinearity is, from Fig. 2,

$$\begin{aligned} f(x_i) &= x_i && \text{for } |x_i| \le V_c \\ &= \Delta(x_i - V_c) + V_c && \text{for } x_i > V_c \\ &= \Delta(x_i + V_c) - V_c && \text{for } x_i < -V_c \end{aligned} \tag{6}$$

By substituting equations 5 and 6 into equation 4 and by evaluating the resulting integrals, the equivalent gain is found to be

$$k_{eq} = \Delta - \frac{2(\Delta-1)}{\sigma\sqrt{2\pi}} \int_0^{V_c} \epsilon^{-x_i^2/2\sigma^2} dx_i \tag{7}$$

This expression is simplified by taking Z as the variable of integration where

$$Z = \frac{x_i}{\sigma\sqrt{2}} \tag{8}$$

and noting that

$$erf(t) = \frac{2}{\sqrt{\pi}} \int_0^t \epsilon^{-t^2} dt \tag{9}$$

The result is

$$k_{eq} = \Delta - (\Delta - 1)\, erf\left(\frac{V_c}{\sigma\sqrt{2}}\right) \tag{10}$$

$$= \Delta\left[1 - \frac{\Delta-1}{\Delta}\, erf\left(\frac{1}{\sigma_n}\right)\right] \tag{11}$$

where σ_n is the normalized input to the nonlinear element.

This normalized gain is plotted in Fig. 3 for two values of Δ.

Signal at Input to Nonlinear Element

For the analysis the signal at the input to the closed-loop system is taken to have a Gaussian amplitude-probability distribution and a spectral density of the form

$$\Phi_{ii} = \Phi_0 \left| \frac{1}{S^2T_f^2 + 2ST_f + 1} \right|^2 \tag{12}$$

where Φ_0 is the zero-frequency spectral density.

The mean-square value of this input is consequently

$$I^2 = \frac{1}{j}\int_{-j\infty}^{j\infty} \Phi_0 \left| \frac{1}{S^2T_f^2 + 2ST_f + 1} \right|^2 dS \tag{13}$$

Integrals of this form are evaluated in reference 10 and are tabulated in the appendix. The solution is, from the appendix,

$$I^2 = \frac{\pi}{2T_f}\, \Phi_0 \tag{14}$$

The quasilinear transfer function from system input to the nonlinear element is the following

$$\frac{g(1+ST_d)}{1 + g(1+ST_d)k_{eq}/ST^2}$$

The mean-square value of the input to the nonlinear element is therefore

$$\sigma^2 = \frac{1}{j}\int_{-j\infty}^{j\infty} \Phi_0 \times \left| \frac{g(1+ST_d)}{[1+g(1+ST_d)k_{eq}/ST^2] \times [ST_f^2 + 2ST_f + 1]} \right|^2 dS \tag{15}$$

It is of advantage, however, to solve this equation in terms of the normalized parameters of the system, which are tabulated in the appendix. Proceeding in this manner it is found that the

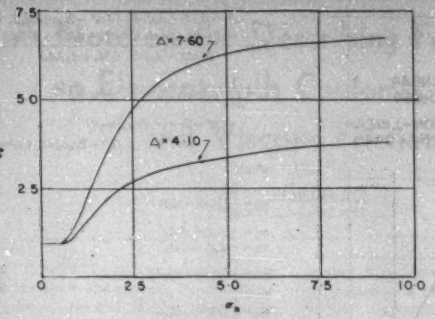


Fig. 3. The equivalent gain loci

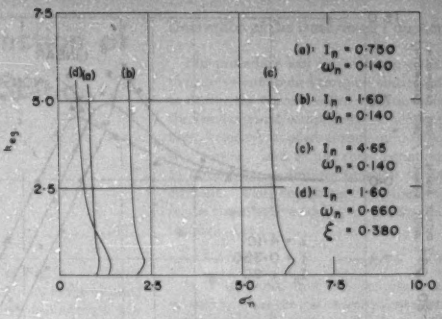


Fig. 4. The input loci

normalized mean-square value of the input to the nonlinear element is

$$\sigma_n^2 = \frac{1}{j}\int_{-j\infty}^{j\infty} \Phi_n \times \left| \frac{(S^2/\omega_0^2 + S^3\xi/\omega_0^3)}{[S^3/\omega_0^3 + Sk_{eq}\xi/\omega_0 + k_{eq}] \times [S^2/\omega_n^2 + 2S\omega_n/\omega_0 + 1]} \right|^2 dS \quad (16)$$

$$= \frac{1}{j}\int_{-j\infty}^{j\infty} \Phi_n \times \left| \frac{(c_2S^2 + c_3S^3)}{d_0 + d_1S + d_2S^2 + d_3S^3 + d_4S^4} \right|^2 dS \quad (17)$$

where

$$\begin{aligned} c_2 &= 1/\omega_0^2 \\ c_3 &= \xi/\omega_0^3 \\ d_0 &= k_{eq} \\ d_1 &= k_{eq}(2\omega_n/\omega_0 + \xi/\omega_0) \\ d_2 &= 1/\omega_0^2 + k_{eq}(2\xi\omega_n/\omega_0^3 + \omega_n^2/\omega_0^2) \\ d_3 &= 2\omega_n/\omega_0^2 + k_{eq}(\xi\omega_n^2/\omega_0^3) \\ d_4 &= \omega_n^2/\omega_0^4 \end{aligned} \quad (18)$$

The solution of equation 17 is, from the appendix,

$$\sigma_n^2 = 2\pi \left\{ \frac{c_2^2[-d_0^2d_3 + d_0d_1d_2] + c_3^2[d_0d_1d_4]}{2d_0d_4[-d_0d_3^2 - d_1^2d_4 + d_1d_2d_3]} \right\} \Phi_n \quad (19)$$

Typical curves are plotted in Fig. 4 for various I_n, the normalized value of the input to the system. These curves are referred to subsequently as the input loci.

Evaluation of Response

Within the closed-loop system the conditions specified by equations 11 and 19 must be satisfied simultaneously. The most convenient way,[8] of determining the conditions for simultaneous solution is to superimpose the k_{eq} loci of Fig. 3 onto the input loci of Fig. 4. Proceeding in this manner it is found that the appropriate loci intersect at only one point. Consequently, for this particular choice of parameters dual-mode operation does not take place. When I_n is small, the input locus lies very near the k_{eq} axis and the equivalent gain of the nonlinear element is approximately unity. The response of the nonlinear system is approximately that for linear operation. As I_n is increased, the input loci move in the direction of increasing σ_n and k_{eq} increases in a corresponding manner. The nonlinear response deviates, therefore, from linear operation. Typical output-input voltage gain spectra have been calculated in this manner and are shown in Fig. 5.

The manner in which the input loci vary with signal bandwidth is shown in Fig. 4. The input loci b and d are for input signals of the same rms value but different bandwidth, the bandwidth of the signal corresponding to b being approximately 4.7 times greater than that corresponding to d. The effect of limiting the signal bandwidth on the output voltage gain-frequency characteristic is shown in Fig. 6. The response denoted by b corresponds to the input locus d and the response denoted by c corresponds to the input locus b.

Experimental Results

The electromechanical system was simulated on an analogue computer and the experimental responses to random signal excitations are shown in Figs. 5 and 6. It is observed that these experimental responses are in good agreement with those calculated.

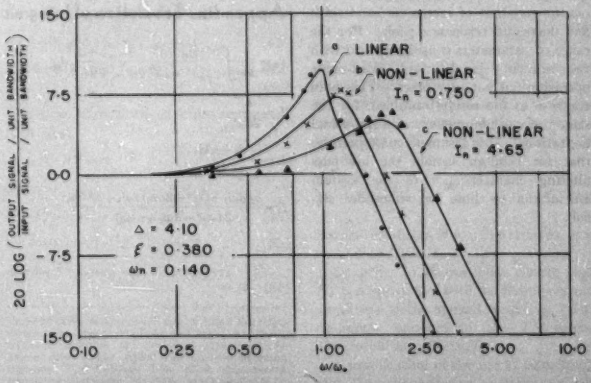


Fig. 5. The response to a random signal showing the effect of varying signal power. Solid line indicates calculated responses. Experimental responses are shown by points

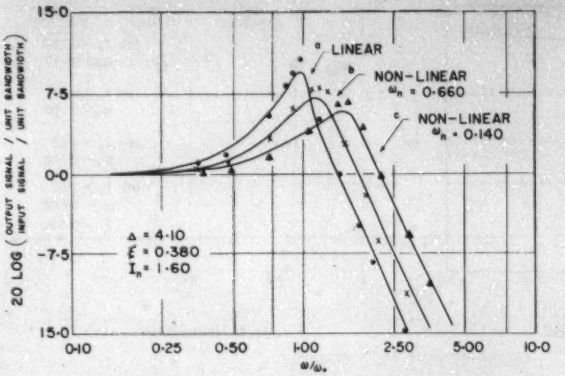


Fig. 6. The response to a random signal showing the effect of varying signal bandwidth. Solid line indicates calculated responses. Experimental responses are shown by points.

Conclusions

From a generalization of the describing function technique, the response of a control system incorporating a hard-spring amplitude nonlinearity has been evaluated for random signal excitations. The signals considered were stationary time series with a Gaussian amplitude-probability distribution and no constant component. It was found that the effect of keeping the signal bandwidth constant and increasing the rms signal amplitude is to produce an output response of increased bandwidth and decreased resonance peak. It was found, also, that the effect of keeping the rms signal amplitude constant and increasing the signal bandwidth is to produce an output response of increased bandwidth and decreased resonance peak. For the range of parameters considered, and with the random signals used, dual-mode operation did not take place. This is in contrast to the results obtained, for the same range of parameters, with sinusoidal excitations. It is assumed, consequently, that for random signals the low-pass filtering characteristics of the system are inferior to those for sinusoidal signals.

It has also been found, although not shown in this paper, that the effect of increasing the signal amplitude and/or bandwidth is to produce error responses of increased bandwidth and decreased resonance peak. This suggests that a hard-spring nonlinearity might be profitably applied to systems in which stability is difficult to achieve without sacrifice of response time.

Finally, the system considered was simulated on an analogue computer and experimental results were obtained. Since there is agreement between the experimental and calculated responses, it is concluded that the technique employed for analysis is reliable.

Appendix. Tabulation of Integrals

$$I_n = \frac{1}{2\pi j}\int_{-j\infty}^{j\infty} \frac{|c_{n-1}S^{n-1}+\ldots+c_0|^2}{d_nS^n+\ldots+d_0}\,dS$$

$$I_1 = \frac{c_0^2}{2d_0d_1}$$

$$I_2 = \frac{c_1^2d_0+c_0^2d_1}{2d_0d_1d_2}$$

$$I_3 = \frac{c_2^2d_0d_1+(c_1^2-2c_0c_2)d_0d_3+c_0^2d_2d_3}{2d_0d_3(-d_0d_2+d_1d_3)}$$

$$I_4 = \frac{c_3^2(-d_0^2d_3+d_0d_1d_2)+(c_2^2-2c_1c_3)d_0d_1d_4+(c_1^2-2c_0c_2)d_0d_3d_4+c_0^2(-d_1d_2^2+d_2d_3d_4)}{2d_0d_4(-d_0d_3^2-d_1^2d_4+d_1d_2d_3)}$$

$$I_5 = \frac{1}{2V_5}[c_4^2m_0+(c_3^2-2c_2c_4)m_1+(c_2^2-2c_1c_3+2c_0c_4)m_2+(c_1^2-2c_0c_2)m_3+c_0^2m_4]$$

where

$$m_0 = \frac{1}{d_0}(d_3m_1-d_4d_2)$$

$$m_1 = -d_0d_4+d_1d_3$$

$$m_2 = -d_0d_3+d_1d_4$$

$$m_3 = \frac{1}{d_0}(d_2m_2-d_3m_1)$$

$$m_4 = \frac{1}{d_0}(d_2m_3-d_3m_2)$$

$$\nabla_5 = d_0(d_1m_4-d_3m_3+d_5m_2)$$

References

1. A Phase-Plane Approach to the Compensation of Saturating Servomechanisms, Arthur M. Hopkin. *AIEE Transactions*, vol. 70, pt. I 1951, pp. 631–39.

2. Non-Linear Techniques for Improving Servo Performance, D. McDonald. *Proceedings*, National Electronics Conference, Chicago, Ill., vol. 6, 1950, pp. 400–21.

3. The Frequency Response of a Servomechanism Designed for Optimum Transient Response, J. C. West, P. N. Nikiforuk. *AIEE Transactions*, vol. 75, pt. II, Sept. 1956, pp. 234–39.

4. Nonlinear Vibrations (book), J. J. Stoker. Inter-Science Publishers, Inc., New York, N. Y., 1950.

5. The Relation Between Force and Velocity in Human Muscle, D. R. Wilkie. *The Journal of Physiology*, London, England, 1950, vol. 110, nos. 3 and 4, p. 429.

6. The Behaviour of a Remote Position Control Servomechanism with Hard-Spring Non-Linear Characteristics, J. C. West, P. N. Nikiforuk. *Proceedings*, Institution of Electrical Engineers, London, England, vol. 101, pt. II, 1954, pp. 481–92.

7. The Response of Remote-Position-Control Systems with Hard-Spring and Non-Linear Characteristics to Step-Function and Random Inputs, J. C. West, P. N. Nikiforuk. *Ibid.*, vol. 102, pt. II, 1955, pp. 575–95.

8. The Analysis of Particular Servomechanisms, P. N. Nikiforuk. *Ph.D. Thesis*, Department of Electrical Engineering, Manchester University, Manchester, England, 1955, pp. 68–91.

9. A Frequency Response Method for Analyzing and Synthesizing Contactor Servomechanisms, Ralph J. Kochenburger. *AIEE Transactions*, vol. 69, pt. I, 1950, pp. 270–84.

10. Non-Linear Control Systems with Statistical Inputs, R. C. Booton. *Report no. 61*, Dynamic Analysis and Control Laboratory, Massachusetts Institute of Technology, Cambridge, Mass.

A Note on the Describing Function of an Element with Coulomb, Static, and Viscous Friction

M. Y. SILBERBERG
ASSOCIATE MEMBER AIEE

IT IS NOW a standard procedure to use the sinusoidal-analysis, or describing-function, method in the investigation of quasilinear servomechanisms. The Nyquist criterion is applied to determine the stability conditions, and the Nyquist diagram provides some information as to the type of equalization that should be employed.[1] These techniques may be used even when the describing-function of the nonlinear element is both frequency and amplitude sensitive, but this requires the graphing of a family of describing functions.[2] This complication may be avoided by isolating the frequency-sensitive and amplitude-sensitive portions of the describing function into separate terms. In the trivial case, the describing function H_0' may be represented as

$$H_0' = H_a H_0 \tag{1}$$

where H_0 is amplitude dependent and frequency independent, and H_a is amplitude independent and frequency dependent. Unfortunately, this easy separation of terms canot be achieved for the describing function of an element with coulomb, static, and viscous friction, a type of element present in many servo systems. This paper will show, however, that an equivalent result is achieved by means of an equivalent, block-diagram representation of H_0'.

Characteristic Curve and Torque Equations

Static friction is defined as the friction torque necessary to initiate motion of a body from rest. Coulomb friction is defined as the friction torque which opposes motion of the body and is a constant, independent of speed. Viscous friction is defined as the friction torque which opposes motion of the body and is proportional to the speed.

The idealized characteristic which describes the situation when these three types of friction are present is shown in Fig. 1.

Consider a rotating body with moment of inertia J and coefficient of viscous friction B, as shown in Fig. 2. τ_f and $\dot{\theta}$ are defined as in Fig. 1. τ_a is the applied torque. On the basis of Newton's law of motion, the torque equations are as follows:

$$\text{for } \dot{\theta}=0^+,\ \tau_a = +\tau_s' \tag{2}$$

$$\text{for } \dot{\theta}>0,\ \tau_a = T_c + B\dot{\theta} + J\ddot{\theta} \tag{2a}$$

$$\text{for } \dot{\theta}=0^-,\ \tau_a = \tau_s' \tag{2b}$$

$$\text{for } \dot{\theta}<0,\ \tau_a = -T_c + B\dot{\theta} + J\ddot{\theta} \tag{2c}$$

In these equations, τ_s' is a variable friction torque which extactly matches τ_a until the magnitude of τ_a exceeds the static torque T_s.

Several other torque quantities will be defined in the following. First, an equivalent applied torque τ_a' such that

$$\tau_a' = \tau_a - B\dot{\theta} \tag{3}$$

Then

$$\text{for } |\dot{\theta}|>0,\ \tau_a' = \pm T_c + J\ddot{\theta} \tag{3a}$$

$$\text{and for } \dot{\theta}=0,\ \tau_a' = \tau_a = \tau_s' \tag{3b}$$

Second, an effective torque τ_e such that

$$\tau_e = \tau_a - \tau_f + B\dot{\theta} \tag{4}$$

Then

$$\text{for } |\dot{\theta}|>0,\ \tau_e = \tau_a \pm T_c = B\dot{\theta} + J\ddot{\theta} \tag{4a}$$

$$\text{for } \dot{\theta}=0, \text{ and } |\tau_a| \le T_s,\ \tau_e = 0 \tag{4b}$$

$$\text{and for } \dot{\theta}=0, \text{ and } |\tau_a| \ge T_s,\ \tau_e = \tau_a - T_s = J\ddot{\theta} \tag{4c}$$

Third, an equivalent effective torque τ_e' such that

$$\tau_e' = \tau_a - B\dot{\theta} = J\ddot{\theta} \tag{5}$$

Then

$$\text{for } |\dot{\theta}|>0,\ \tau_e' = J\ddot{\theta} = \tau_a' \pm T_c \tag{5a}$$

$$\text{for } \dot{\theta}=0, \text{ and } |\tau_a| \le T_s,\ \tau_e' = 0 \tag{5b}$$

$$\text{and for } \dot{\theta}=0, \text{ and } |\tau_a| \ge T_s,\ \tau_e' = \tau_a' - T_s = J\ddot{\theta} \tag{5c}$$

Paper 56-960, recommended by the AIEE Feedback Control Systems Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Fall General Meeting, Chicago, Ill., October 1-5, 1956. Manuscript submitted September 24, 1954; made available for printing July 24, 1956.

M. Y. Silberberg is with Yale University, New Haven, Conn.

Derivation of the Describing Function

The procedure will be to assume that the equivalent applied torque is sinusoidal; to derive the corresponding equations for the fundamental components of acceleration, velocity, applied torque, and effective torque; and to use these equations to obtain the describing function of the element. The describing function will relate the effective torque to the applied torque. Let

$$\tau_a' = T_a' \sin \omega t \tag{6}$$

It is apparent, from the definitions of τ_a' and τ_s', that the relationship that exists between τ_a' and τ_s' is exactly that which exists between τ_a and τ_s when the coefficient of viscous friction is zero. The latter relationship has already been obtained.[3,4] The fundamental component of τ_s' is therefore given by

$$\tau_s' = CT_a' \sin(\omega t + \delta) \tag{7}$$

where C and δ are functions of T_a', T_s, and T_c.[3,4] τ_e and τ_a are obtained by using this result and the definitions given in equations 2 to 6.

$$\tau_e = CT_a' \sin(\omega t + \delta) - \frac{BCT_a'}{\omega J} \cos(\omega t + \delta) \tag{8}$$

$$\tau_a = T_a' \sin \omega t - \frac{BCT_a'}{\omega J} \cos(\omega t + \delta) \tag{9}$$

The describing function, which is to relate the effective torque to the applied torque, is defined as the ratio of the Laplace transforms of τ_e and τ_a evaluated for $s=j\omega$.[1] Performing the transformations and obtaining this ratio yields

$$H_0' = \frac{(1+TS)C(\cos\delta + j\sin\delta)}{TS + C(\cos\delta + j\sin\delta)} = \frac{(1+TS)C/\underline{\delta}}{TS + C/\underline{\delta}} \tag{10}$$

where $T = J/B$. But $C/\underline{\delta}$ is what has previously been defined as H_0, the describing function relating τ_e and τ_a when the coefficient of viscous friction is zero.[3,4] Substituting $H_0 = C/\underline{\delta}$ into equation 10 yields

$$H_0' = H_0 \frac{(1+TS)}{(H_0+TS)} \tag{11}$$

For the condition $B=0$, this reduces, as it should, to $H_0' = H_0$.

Equation 11 shows quite clearly that H_0' is a function of both the frequency and amplitude of the applied torque τ_a. The frequency appears in the TS terms. The amplitude affects H_0.

A justification of the use of equation 6 and some background material on the significance of the describing function H_0 are given in Appendixes I and II.

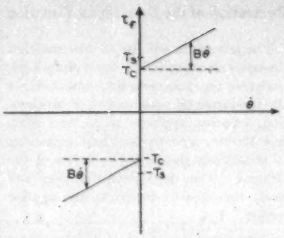
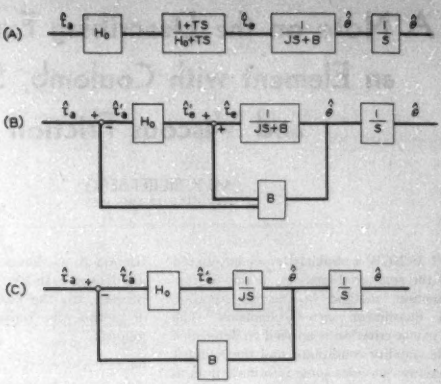


Fig. 1. Characteristic curve for an element with coulomb, static, and viscous friction. τ_f is the total friction torque, T_s is the magnitude of the static friction torque, T_c is the magnitude of the coulomb torque, B is the proportionality constant relating speed and viscous torque, and $\dot{\theta}$ is the angular velocity of the body

Fig. 3. Block-diagram representation of H_0'



Alternative-Derivation Block-Diagram Representation of H_0'

Equation 11 may be derived in a slightly different manner. This derivation will lead to a useful block-diagram representation of H_0' which isolates the amplitude-sensitive and frequency-sensitive factors into a forward path, and a feedback path, respectively.

On the basis of ideas already presented, it is possible immediately to write a frequency domain equation relating τ_e' and τ_a'. This equation is

$$\hat{\tau}_e' = H_0 \hat{\tau}_a' \tag{12}$$

The other torque equations may also be written in the frequency domain. These are

$$\hat{\tau}_a' = \hat{\tau}_a - B\hat{\theta} \tag{13}$$

$$\hat{\tau}_e = B\hat{\theta} + J\hat{\theta} = (B+JS)\hat{\theta} \tag{14}$$

$$\hat{\tau}_s = J\hat{\theta} = JS\hat{\theta} \tag{15}$$

Combining these equations,

$$\frac{\hat{\tau}_s}{\hat{\tau}_a} = H_0' = H_0 \frac{(1+TS)}{(H_0+TS)} \tag{16}$$

which checks the result obtained previously.

The equations can also be used to draw block diagrams which represent the transfer between $\hat{\tau}$ and $\hat{\theta}$. Thes eare shown in Fig. 3. The diagram in Fig. 3(C) will undoubtedly be the most useful in the analysis of servo systems.

The frequency-sensitive and amplitude-sensitive portions of the H_0' transfer have therefore been isolated by means of equivalent block diagrams. The amplitude-sensitive portion appears as H_0, a describing function which is known and has already been drawn in the Nyquist plane.

Application of the Block Diagram

Consider the servo system represented by the block diagram of Fig. 4. If the H_0' function is used directly, the negative loop gain function is

$$A = H_1 H_0' \left[\frac{1}{S(JS+B)}\right] \tag{17}$$

The stability of the system is then determined by plotting

$$H = \frac{H}{S(JS+B)} \tag{18}$$

and the critical loci $-1/H_0'$, a different plot of $-1/H_0'$ being needed for each frequency that is considered.

If the equivalent representation is used, only one critical locus is necessary. The block diagram is shown in Fig. 5. Both loops are broken simultaneously, thereby satisfying the conditions for direct application of the Nyquist criterion.[5] The negative loop gain in this case is the following equation

$$A_e = H_0 \frac{1}{JS}\left[B + \frac{H_1}{S}\right] = H_0\left\{\frac{1}{TS} + \frac{H_1}{JS^2}\right\} \tag{19}$$

The stability of the system can be determined by plotting

$$H = \frac{1}{TS} + \frac{H_1}{JS^2} \tag{20}$$

and the single critical locus, $-1/H_0$. This procedure is illustrated in Fig. 6. It is the "stiction locus" that is used here, as explained in Appendix II. Two cases are shown. In one

$$H_1 = K \tag{21}$$

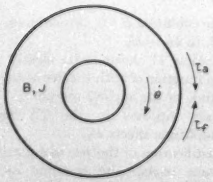


Fig. 2. Rotating body with inertia and friction

Fig. 4. Block diagram of a typical servo

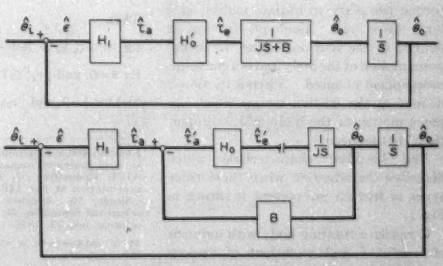


Fig. 5. Equivalent representation of the circuit of Fig. 4

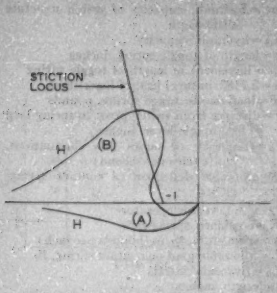


Fig. 6. Illustrating the superposition of the $-1/H_0$ locus and the H plot for the system of Fig. 5

A—$H_1 = K$, a stable system
B—$H_1 = (1+T_1S)/(1+T_2S)$, an unstable system

a constant, and the system is stable. In the other

$$H_1 = \frac{1+TS}{1+T_2S} \tag{22}$$

which, for a certain relationship between the various constants, results in the system shown in Fig. 6. This system is stable for very small inputs but reaches a point of convergent equilibrium for larger inputs. Oscillations are sustained at the frequency and amplitude determined by the intersection point near the origin.

Appendix I. Justification of the Use of Equation 6

The basic assumption that is made in the application of the describing function method is as follows:

Given:
(*a*). A nonlinear element for which the ratio of output to input may depend on both the frequency and amplitude of the input.
(*b*). A sinusoidal input to this element.

Then:
The output of the element can be represented, to a first approximation, by the fundamental component of the Fourier series of the output.

The describing function is the ratio of the Laplace transforms of the fundamental component of the output and the sinusoidal input evaluated at $s = j\omega$. Extending the basic assumption, it is assumed that the describing function is very nearly the transfer function of the element.

These assumptions are valid under very general conditions, particularly in servo systems where low-pass filters attenuate the high-order harmonics. It has been shown that the error introduced by neglecting the effect of these harmonics is no more than 5 to 10 per cent for most servo systems.[2]

It has also been shown that the assumptions are valid for the particular case of an element with coulomb and static friction.[3,4] It is therefore proper to relate τ_0' and τ_0' by equations 6 and 7, and furthermore, as a consequence of equations 2 through 5, to consider only the fundamental components of $\dot{\delta}$, δ, τ_0, and τ_s.

Appendix II. The Describing Function H_0

The following material describes briefly the manner in which the describing function H_0 was obtained. This information is derived from references 3 and 4, and is presented here for the sake of clarity.

An explanation of the basic assumptions used in the derivation has been given in Appendix I. With the coefficient of viscous friction set equal to zero, τ_0 was assumed to be sinusoidal. The fundamental component of τ_0 was obtained by considering the steady-state wave forms of $\dot{\delta}$, δ, and τ_s. Consideration was given to the case when $\dot{\delta}$ had a dead zone, as well as to the case when $\dot{\delta}$ did not have a dead zone. Equations for the amplitude, C, and phase, δ, of the describing function were then obtained in the manner discussed in Appendix I. The resulting equations involved such terms as λ, α_1, α_2, and β, where

$$\begin{aligned} \lambda &= T_s/T_0 \\ \alpha_1 &= \sin^{-1}\lambda \\ \alpha_2 &= \sin^{-1} T_s/T_0 \\ \beta &= \cos^{-1}(\pi\lambda/2) \end{aligned} \tag{23}$$

It is apparent that a polar plot of $-1/H_0$ results in a family of curves. One possible family, the static friction loci described in the reference, are plots of $-1/H_0$ for various ratios of static to coulomb friction. These loci are shown in Fig. 7. The dotted line which joins the end points of these loci has been defined as the stiction or S locus.

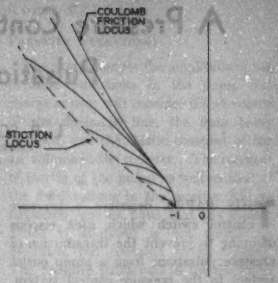


Fig. 7. Static friction loci, and S locus

The stability criterion for closed-loop systems that are open-loop stable can be expressed in terms of the relative positions of these loci and a plot of H_1 where the open-loop gain is given by

$$A = H H_1 \tag{24}$$

The system will be absolutely stable as long as the $-1/H_0$ locus lies to the left of the H_1 locus.[1]

If the ratio of static to sliding friction is known for a particular system, it is only necessary to plot the static friction locus corresponding to that ratio. A generally more conservative design, but one which will be applicable for all ratios of static to sliding friction can be based on the stiction locus.

References

1. A Frequency Response Method for Analyzing and Synthesizing Contactor Servomechanisms, Ralph J. Kochenburger. *AIEE Transactions*, vol. 69, pt. I, 1950, pp. 270–84.

2. Sinusoidal Analysis of Feedback-Control Systems Containing Nonlinear Elements, E. Calvin Johnson. *Ibid.*, vol. 71, pt. II, July 1952, pp. 169–81.

3. Coulomb and Static Friction in Servomechanisms, Julius Tou. *Ph. D. Thesis*, Electrical Engineering Department, Yale University, New Haven, Conn., 1953.

4. Static and Sliding Friction in Feedback Systems, J. Tou, P. M. Schultheiss. *Journal of Applied Physics*, New York, N. Y., vol. 24, no. 9, Sept. 1953, p. 1210.

5. Theory of Servomechanisms (book), H. M. James, N. B. Nichols, R. S. Phillips. McGraw-Hill Book Company, Inc., New York, N. Y., 1947, chap. 2.

A Pressure Control Switch with a Pulsation Filter

L. P. SCHAEFER
MEMBER AIEE

THIS PAPER describes a pressure control switch which uses viscous damping to prevent the transmission of pressure pulsations from a pump outlet system to the pressure control system. This was accomplished by designing a pressure switch which has certain dynamic characteristics as well as the normal static characteristics. By controlling the amplitude of the pressure pulsations at the switch with a mechanical low-pass filter, a resulting simplification of the switch structure was obtained.

Several devices have been in use to reduce pressure pulsations, such as the ball-type check valve and small-orifice plugs. These devices have aided in reducing pressure peaks by an amount relative to the size of the opening through them. Since the size of the opening in these devices had to be kept large enough to prevent clogging with foreign matter, and to ensure proper operation, the resultant pulsation reduction was not very great.

Various methods of improving the pressure-switch dynamic performance were investigated and this resulted in the control switch to be described. The control switch is found to be capable of reducing pressure pulsations by 90 per cent or better. The steady-state phase angle between the pump pulsations and the reduced pulsation at the switch was approximately 85 degrees. The paper shows how this was accomplished. The switch is shown in Fig. 1.

A 2-pole single-throw pressure control switch with a viscous damper was developed for use in pulsating pressure systems. The switch was designed to operate, without pulsating action, in a 2-piston pump discharge system, where the pump speed is 210 rpm, or greater. The switch, however, may be readily modified to operate in a system other than that just described, and the factors affecting these design constants will be pointed out.

The switch was designed to have a static pressure range of 0 to 60 psi (pounds per square inch) and a pressure differential of 13 to 40 psi. The switch will control a 1-horsepower electric motor load, for 110 to 220 volts, 1- or 3-phase. The viscous damper uses silicone fluid, resulting in a small temperature change in the damper characteristic. The steel constructed switch has the over-all dimensions of $4^1/_2$ inches long by $2^{13}/_{16}$ inches wide by 3 inches high.

Control Switch and Pump System

A typical piston type of pump with the control switch installed in the system is shown in Fig. 2. This is a simple, automatic, closed-cycle type of pressure control system. When the pressure switch contact is closed, the electric motor is connected to the supply voltage. The static pressure in the pump discharge system increases, as the pump operates, until a "set" value of pressure is reached which opens the control switch contacts. This stops the motor and the pump, and the pressure gradually decreases. The pressure switch contacts then make at a set value of pressure and the cycle is repeated again.

On the pressure increase cycle, with the pump operating, the pressure increases in a pulsating fashion. The value of the amplitude of this pulsation varies, depending on where it is measured. It is greatest in the pump cylinder, less in the discharge line, and hardly measureable at the air volume of the storage tank. When the pressure switch is connected to the discharge line, and without a pulsation damper, the switch-operating structure will vibrate in unison with the pressure pulsations. Before, and at, the point of contact break, the switch will give erratic control by stopping and starting the pump motor. This is objectionable from an operation standpoint, as well as that of switch and motor life.

Nomenclature

A = constant, differential equation notation
A_s = effective area of diaphragm, square inches
A_p = piston area, square inches
a = lever arm from switch pivot to spring, inches
b = lever arm from switch pivot to pressure point, inches
C = hydraulic capacity of switch front diaphragm
$_mC$ = hydraulic capacity of switch structure diaphragm
C_h = hydraulic capacity
c = length of toggle spring, inches
dc = increment of length of toggle spring
e = 2.7183, natural base of logarithms
F = load due to range spring, pounds
f = distance from switch pivot to spring loop on toggle lever, inches
f_p = frequency of pump flow fluctuations, cps (cycles per second)
g = allowable deflection of contact spring, inches
$j = \sqrt{-1}$, vector operator
K = combined spring rate
k = spring rate, lb/in (pounds per inch)
L = allowable load on contact spring, lb
L_h = hydraulic inertia
l = length, inches
n = number, 0.25 or 0.18, for switch
P = pressure
P_0 = pressure P_1 at time $t = 0$
P_1 = pressure at switch inlet
P_2 = pressure at switch strueture
P_a = absolute pressure
Q = amplitude of flow fluctuation
q = instantaneous flow
R = hydraulic resistance of switch viscous damper
R_h = hydraulic resistance
r = radius, inches
rpm = pump speed, rpm
S = load, pounds, at switch pressure point
s = specific gravity of the fluid
t = time, sec (seconds)
V = volume, in^3
V_p = maximum piston velocity, in/sec
v = volume displacement
v_p = instantaneous piston velocity
x = travel at pressure point, inches
dx = increment of travel at pressure point
y = travel at range spring, inches
α = constant, differential equation notation
β, γ = angle indicated in Fig. 12
η = absolute viscosity, poises
θ = angle indicated in Fig. 12
θ_p = shaft angle from position where piston 1 is about to begin discharge stroke
π = 3.1416
ϕ = angle indicated in Fig. 12
ω_p = frequency of pump flow fluctuations, rad/sec (radians per second)
ω = same as ω_p

The electrical analogy system used in this paper is shown in Table I.

Table I. Analogy System and Symbols Used

Symbol	Pressure	Electrical
p	lb/in^2	E, volt
v	in^3	Q, coulomb
q	in^3/sec	I, ampere
C_h	in^3/lb/in^2	C, farad
R_h	lb/in^2-sec/in^3	R, ohm
L_h	lb/in^2-sec^2/in^3	L, henry

Paper 56-932, recommended by the AIEE Industrial Control Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Fall General Meeting, Chicago, Ill., October 1-5, 1956. Manuscript submitted June 11, 1956; made available for printing August 9, 1956.

L. P. SCHAEFER, formerly with The Hinchman Corporation, Detroit, Mich., is now with the General Electric Company, Bloomington, Ill.

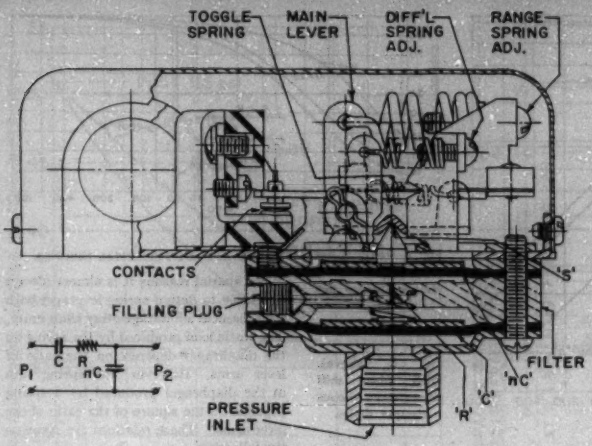


Fig. 1. Pressure control switch

In insert
P_1 = pressure in the pump discharge line
P_2 = pressure across switching mechanism

Pump Discharge System

To properly design a pressure pulsation damper, it is necessary to analyze the piston-pump discharge system. Such an analysis for hydraulic equipment has been described.[1]

Pump Flow Fluctuations

The motion of piston-type pumps is sinusoidal, which may be expressed thus

$$v_p = V_p \sin \theta_p$$

Then the discharge rate of a piston is the area of the piston times its velocity, and has been given as

$$v_p A_p = V_p A_p \sin \theta_p = Q \sin \theta_p$$

The flow fluctuation of a double-acting piston pump is shown in Fig. 3. Every other peak is shorter due to the piston area reduction by the piston rod. For purposes of calculation, a sinusoidally varying flow is assumed for convenience, rather than the "scallop" form shown in Fig. 3. Little error is introduced when the amplitude of the flow fluctuation is made two-thirds of the amplitude shown in Fig. 3. By the Fourier method the lowest frequency of the pump flow pulsations is seen to be twice the pump frequency; however, a factor 1.5 is used.

Frequency of Flow Fluctuations

The forcing frequency of the pump flow pulsations is given as

$$f_p = \text{rpm}/60 \times 1.5 \text{ cps}$$

Pressure Pulsations

The expression for the amplitude of the pressure pulsations in the pump discharge system will consider the geometry of the discharge line, the fluid being pumped, and the elastic constant of the air volume in the receiver. The moment of inertia of the pump as well as that of the fluid volume in the receiver will be neglected, on the basis that the system which is analyzed is not completely hydraulic but contains an air volume. The pressure pulsations in the discharge pipe and the air volume of the receiver are given by

$$P = qj\omega_p L_a + qR_a + q/j\omega_p C_h$$

This expression for pressure pulsations is derived, and a sample design calculation is given in Appendix I. For pump *A*, which was calculated, it was found that the amplitude of the pressure pulsation in the discharge pipe was 5.34 psi at a frequency of 47.5 rad/sec. A second pump was calculated, the results of which are also in Appendix I.

Viscous Damping Device

The pressure control switch is described in two parts: the viscous damper and the switch structure.

Filter Investigation

An *RC* type of electric network has been used as the basis of the pulsation damper design. This is covered in the literature[2] on equalizers, etc. Another reason for this approach is that the mechanical analogies of the *RC* electrical components are easily reproduceable. The two points of interest in the filter design are that the static pressure must be transmitted across the filter, and pulsations of the frequency of the pump must not pass through the filter. The simplest form of

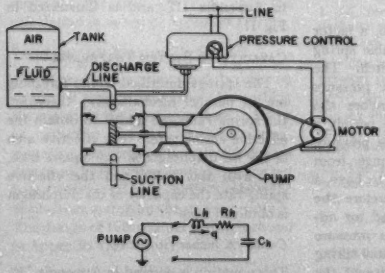


Fig. 2. Control switch and pump system

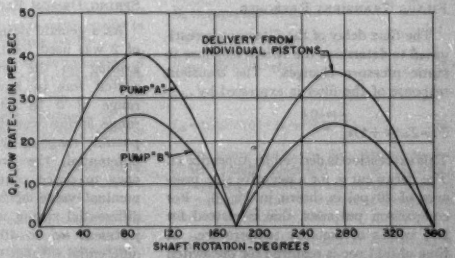


Fig. 3. Pump flow fluctuations

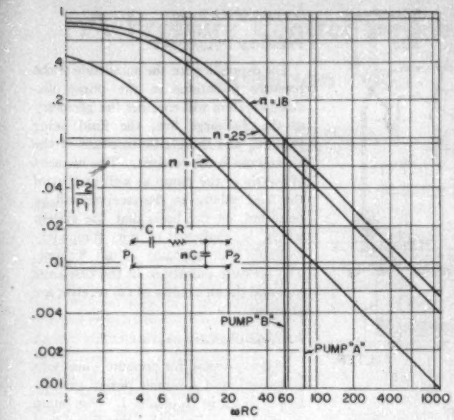


Fig. 4 (left). Filter pressure-gain curves, viscous damper design

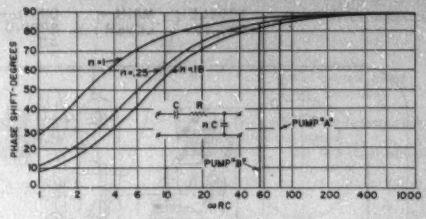


Fig. 5 (above). Filter phase shift curves, viscous damper design

this low-pass filter is a resistance and capacitance in series. The resistance, mechanically, is a fluid of a particular viscosity flowing through an orifice of certain dimensions. The viscous fluid is contained between two resilient diaphragms. The diaphragms with their respective spring systems form, mechanically, the capacitor part of the filter.

Filter Steady-State Response

The low-pass filter is represented electrically, together with the component parts, in Fig. 1. The following relation is obtained for the pressure-gain across this network

$$P_2/P_1 = 1/(1+n+j\omega_p nRC)$$

This expression is derived in Appendix II and shown in Fig. 4. The phase shift for the filter is shown in Fig. 5. At the pump frequency P_2 is 85 degrees out of phase with P_1.

Filter Transient Response

The time delay of the filter was investigated to determine the switch response to static pressure changes. The transient response of the filter is expressed by

$$P_2 = P_0/(n+1)\epsilon^{\frac{-(n+1)\,t}{nRC}}$$

This expression is derived in Appendix II. The decay curve for a negative step pressure of 20 psi is shown in Fig. 6. For comparison purposes, this is plotted for two values of ambient temperature. A time of 0.14 sec is required for the switch pressure to drop 20 psi at 77 F (degrees Fahrenheit). This is a satisfactory value.

Pressure Control Switch Structure

The first consideration in the switch structure design is that it meet the requirements for static operation of the switch. Second, by the contribution the structure spring loading to the nC diaphragm, the switch structure becomes part of the viscous damper and contributes to the dynamic characteristic of the switch.

Static Pressure Division Across Capacities

The static pressure division for P_1 and P_2 due to the spring loads on the two diaphragms C and nC is given as

$$P_2 = P_1/(1+n)$$

This relation is used to refer the nC diaphragm load pressures to the switch pressure inlet.

Spring Design Factors

As a general design criterion, a factor of 2 was used for the adjustable spring loading on the pressure switch. The switch was to have a static pressure range of 0 to 30 psi; therefore the range spring was designed to permit a possible range of 60 psi. The pressure differential, the pressure change from open to closed contacts, was to have a nominal value of 20 psi; therefore the differential spring was designed for adjustment up to 40 psi. The pressure differential without the differential spring in the system, is the sensitivity of the switch. This value is 13 psi.

Range and Differential Springs

For spatial reasons it is almost always necessary to design spring leverages with a mechanical advantage other than unity. The static load produced by the spring on the diaphragm depends on the ratio of lever arms. However, the spring rate at the diaphragm produced by a spring depends on the square of the ratio of the lever arms. These relations are given as the following

$$S = Fa/b$$

and $$S/x = k(a/b)^2$$

These equations are derived in Appendix III, and the rates for the range and differential springs are shown in Fig 7.

Toggle Spring

The toggle spring is an overcenter operating spring which causes the contact lever to snap up or down. The mechanics of toggle systems have been described in the past.[3] The toggle spring for this pressure switch contributes very little to the static range of the switch. However, the toggle spring rate will add to the dynamic design of the nC diaphragm. The spring rate of the toggle spring at the diaphragm is given as the equation

$$S/dx = \frac{k(af/b\,\sin\theta)^2}{a^2+f^2-2af\cos\theta}$$

This relation is nonlinear. It is derived in Appendix III, and is illustrated in Fig 7.

Capacity of Switch Diaphragm

The springs described in the foregoing are in parallel mechanically; therefore, the spring rates are added to obtain the effective spring rate. The effective area of the nC diaphragm is 1.23 square inch, and with this value and the effective spring rate the capacity of the diaphragm is obtained from

$$C = A_e^2/K \text{ inches}^3/\text{lb/inch}^2$$

This equation is derived in Appendix III, and the capacity is plotted in Fig. 7.

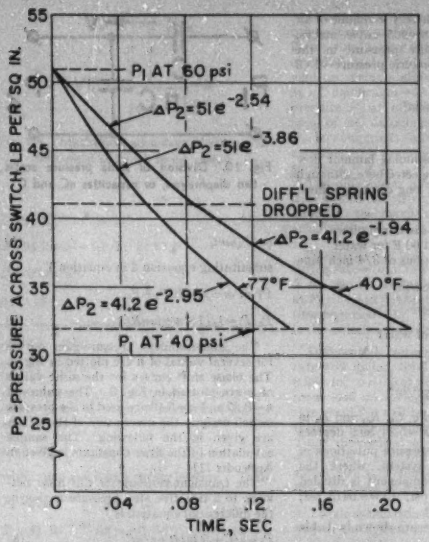


Fig. 6. Filter transient response. Pressure decay across nC diaphragm versus time, for a negative step pressure of 20 psi at the switch inlet. Viscous damper design, R = 94 lb/in³ − sec/in³ at 77 F

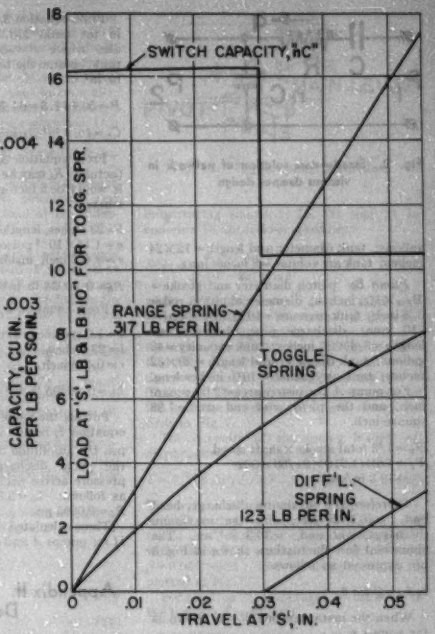

SWITCH CAPACITY, "nC"
RANGE SPRING 317 LB PER IN.
TOGGLE SPRING
DIFF'L. SPRING 123 LB PER IN.
LOAD AT 'S', LB & LB = 10" FOR TOGG. SPR.
CAPACITY, CU IN. PER LB PER SQ IN.
TRAVEL AT 'S', IN.

Fig. 7 (right). Switch design curves, spring loading and capacity

Capacity of the Front Diaphragm

Knowing the value of nC, the value of C of 0.01805 cubic inch/lb/square inch for the front, or inlet pressure, diaphragm is obtained. This is a relatively simple procedure and is shown in Appendix III.

Resistance of the Viscous Damper

The resistance element of the filter is designed by a choice of silicone fluid of a particular viscosity grade, and the diameter of the orifice through which the fluid passes. This value is 94 lb/square inch-sec/cubic inch at 77 F, and the calculation is shown in Appendix III.

The Contact Follow-Up Spring

Butting-type contacts will bounce on striking if there is no means of absorbing the energy of the moving system. A way of reducing the bounce by spring-mounting the stationary contact has been described.[4] On the switch described here, however, it was easier to place the antibounce spring on the moving contact. The design of this contact spring is shown in Appendix III.

Toggle Lever Counterweights

The toggle lever is counterweighted so that the pressure switch will operate the same in any orthogonal position; and also so that the effective mass of the moving contact system will be reduced to a very low value. The design value of the counterweights was determined by taking horizontal and vertical moments about the knife-edge pivots of the toggle lever. This determination is a routine calculation and will not be described in this paper.

Conclusions

1. A pressure control switch has been designed which will operate satisfactorily in any given pressure pulsating system.

2. The amount of amplitude reduction of pressure peaks is 90 per cent for this switch under the conditions described.

3. As the temperature decreases, or the speed of the pump increases, the reduction of pressure peaks will be greater.

4. The switch has an adjustable static pressure range of 0 to 60 psi, and an adjustable pressure differential of 13 to 40 psi.

5. The switch has a counterbalanced contact lever, and may be mounted in any position without affecting the operation.

6. The switch has a 2-pole contact structure with electrical spacings as specified by the Underwriters' Laboratories.

Appendix I. Calculation of Pump Discharge Pressure

The pump discharge system was analyzed by using the electrical analogy[4] system defined in Table I. The pump discharge system is shown in Fig. 2, from which the following equation is written

$$P = qj\omega_p L_3 + qR_3 + q/j\omega_p C_3 \qquad (1)$$

where

$$C_3 = V/P_a \qquad (2)$$

and

P_a is expressed in lb/in²

For the discharge pipe

$$R_3 = 3.69 \times 10^{-4} \eta l/r^4 \qquad (3)$$

also

$$L_3 = 2.59 \times 10^{-2}\, sl/\pi r^2 \qquad (4)$$

The pump shown in Fig. 2 may be for either pump described in the following paragraphs. A sample calculation is given for pump A, while the results of the calculation for pump B are indicated.

Pump A: piston diameter and stroke = $1^5/_8 \times 1^1/_{32}$ inches; diameter of piston rod = 1/2 inch; tank pressure = 50 lb/in²; speed = 302 rpm; discharge pipe diameter and length = 3/4 × 23 inches; tank capacity = 12

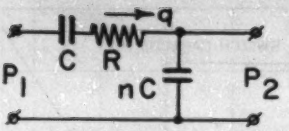


Fig. 8. Steady-state solution of network in viscous damper design

gallons; tank diameter and length = 12×24 inches; tank air volume = 8 inches long.

Pump B: piston diameter and stroke = $1^{5}/_{16}\times1^{1}/_{4}$ inches; diameter of piston rod = 3/8 inch; tank pressure = 40 lb/in²; speed = 210 rpm; discharge pipe diameter and length = 3/4×25 inches; tank capacity = 42 gallons; tank diameter and length = 20×32 inches; tank air volume = $10^{1}/_{2}$ inches long.

For pump *A* the piston area = 2.08 square inch, and the piston rod end area = 1.88 square inch.

V_p = 1/2 total stroke × shaft speed

$V_p = 1/2\times1.219\times2\pi/60\times302$

$= 19.3$ in/sec

Therefore the maximum discharge, head end, = 40.1 in³/sec, and the maximum discharge, rod end, = 36.3 in³/sec. The sinusoidal flow fluctuations shown in Fig. 3 are expressed as follows

$v_p A_p = Q \sin\theta$

When the instantaneous flow is shown as the equation

$q = 2/3Q \sin\omega_p t$

where

$f_p = rpm/60\times1.5 = 7.55$ cps

$\omega_p = 2\pi\times7.55 = 47.5$ rad/sec

Therefore

$q = 28 \sin 47.5t$

From equation 2, where V = volume of air in the tank, 113.2×8 = 905 cubic inches, also where P = absolute pressure in the tank, assume the barometric pressure = 14.3 lb/in², or

$P = 50 + 14.3 = 64.3$ lb/in²

$C_h = 14.2$ in³/lb/in²

From equation 3, assuming laminar flow (actually R_h may be neglected here, although it would be a factor in long discharge lines) where

l = 23 inches, length of discharge pipe
$\eta = 1.15\times10^{-5}$ poises at 60 F for water
r = 0.412 inch, inside radius of 3/4-inch pipe

$R_h = 0.00034$ lb/in²−sec/in³

From equation 4 where

s = 1.0 specific gravity of water
l = 23 inches
r = 0.412 inch

$L_h = 0.00405$ lb/in²−sec²/in³

Putting the values of q, C_h, R_h, and L_h in equation 1, results in $P = 5.34/89.9$ degrees psi, the amplitude of pressure pulsations in the pump discharge system, where the pressure across each component is divided as follows: L_h = 5.37 psi; C_h = 0.0415 psi; R_h = 0.0095 psi.

These calculated data are shown in Table II for pumps *A* and *B*.

Appendix II. Viscous Damper Design

The steady-state solution of the network is made to determine the gain, or the ratio of pressure "out" to pressure "in," and the phase shift; see Fig. 8.

For the network, if

$P_1 = p_1 \sin\omega t$

then

$P_1 = q/j\omega C + qR + q/j\omega nC$ (5)

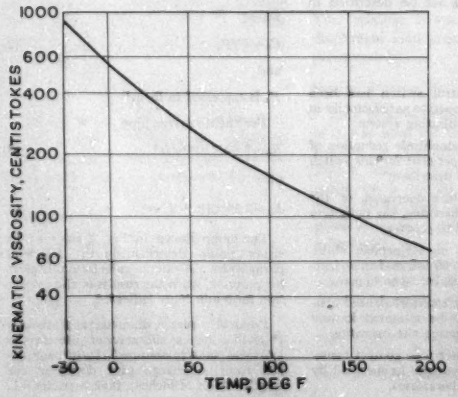


Fig. 9. Viscosity-temperature curve for Dow Corning type of 200 fluid, d-c 200 centistokes, at 77 F, s = 0.971; viscous damper design

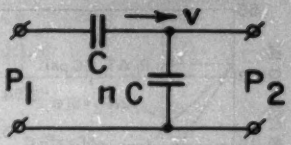


Fig. 10. Division of static pressure across two diaphragms, or capacities nC and C

$q = P_2 j\omega nC$ (6)

substituting equation 2 in equation 1

$P_1 = P_2 n + j\omega nRCP_2 + P_2$

$P_2/P_1 = 1/(1+n+j\omega nRC)$ (7)

The dimensionless pressure-gain curves for several values of n are plotted in Fig. 4. The phase shift curves for the same values of n are plotted in Fig. 5. The values of n = 0.25 and n = 0.18 are used in the pressure switch design; the description of the values are given in the following. The sample calculation of the filter constants is given in Appendix III.

The transient response of this filter network to a negative step pressure is shown; the differential equation is

$P_1 = v/C + Rdv/dt + v/nC$ (8)

where

$q = dv/dt$
$v = nCP_2$

$dv/dt = nCdP_2/dt$

substituting in equation 8

$P_1 = (n+1)P_2 + nRCdP_2/dt$ (9)

let $P_2 = Ae^{-\alpha t}$

then $dP_2/dt = -\alpha Ae^{-\alpha t}$

Substituting in equation 9

$P_1 = (n+1)Ae^{-\alpha t} - nRC\alpha Ae^{-\alpha t}$

$= [(n+1) - nRC\alpha]Ae^{-\alpha t}$

set $(n+1) - nRC\alpha = 0$

$\alpha = (n+1)/nRC$

then $P_2 = Ae^{\frac{-(n+1)}{nRC}t}$

When $t = 0$, $P_2 = A$ and $dv/dt = 0$, and from equation 8,

$P_0 = (1/C + 1/nC)nCP_2 + 0$

From this, then

$P_2 = P_0/(n+1) = A$

therefore

$P_2 = P_0/(n+1)e^{\frac{-(n+1)}{nRC}t}$ (10)

The transient response of P_2 from equation 10 is shown in Fig. 6. This is for a step pressure drop of 20 psi at P_1, or in the pump discharge system, when the static pressure is suddenly reduced from 60 to 40 psi. The following constants from Appendix III, are substituted in equation 10.

Fig. 11. Mechanical advantage of range and differential of switch spring system

$C = 0.01805$ in³/lb/in²

$R = 94$ lb/in²−sec/in³ at 77 F

$P_0 = 60$ lb/in²

$n = 0.18$

with range, differential, and toggle spring loading. Also

$P_3 = 51.5$ lb/in²

which is the value of P_1 where the differential spring is dropped from the loading. Also

$n = 0.25$

with range and toggle spring loading. Then

$P_3 = 51e^{-1.8t}$ for $n = 0.18$

$P_3 = 41.2e^{-1.8t}$ for $n = 0.25$

These last two equations are shown in Fig. 6. Similar curves are also shown for an ambient temperature of 40 F. The temperature-viscosity relation for Dow-Corning silicone fluid, type *DC 200*, with viscosity grade of 200 centistokes at 25 degrees centigrade, is shown in Fig. 9.

Appendix III. Control Switch Design

In this section the design calculations of the switch constants are shown including the viscous damper design. The division of static pressure across the two diaphragms, or capacities nC and C, becomes as shown in Fig. 10.

$$P_1 = v/C + v/nC$$

$$v = P_3 nC$$

$$P_3 = \frac{CP_1}{C+nC} = P_1/(1+n) \qquad (11)$$

The value of R is not considered because it depends on dv/dt.

The switch spring system includes the range and differential which have the mechanical advantage shown in Fig. 11.

$$S = Fa/b \qquad (12)$$

also $F = ky$ (k = spring rate) (13)

and $x = yb/a$ (14)

Substituting equations 12 and 14 in 13

$$Sb/a = kxa/b$$

$$S/x = k(a/b)^2 \qquad (15)$$

where S = resulting load at diaphragm, lb.

The range spring calculation (the spring constants) is as follows: spring rate = 79.2 lb/in; total spring adjustment = 0.375 inch; $a = 0.750$ inch, $b = 0.375$ inch; maximum load on spring = adjustment × rate; load S at nC diaphragm = 60 lb (use equation 12); pressure P_3 at nC diaphragm = S/effective area of nC diaphragm (use equation 11); P_1 is computed; spring rate at S from equation 15 is determined; $S/x = 317$ lb/in, shown in Fig. 7.

The differential spring calculation, where the switch sensitivity is calculated and added to the maximum load of the differential spring, is as follows:

Sensitivity: range spring rate + toggle spring rate = total rate; load at S = travel at S to operate switch × total rate; pressure at P_3 = load at S/effective area of diaphragm (use equation 11); $P_1 = 13.1$ lb/in², which is the switch sensitivity.

Differential spring constants: spring rate = 69 lb/in; total spring adjustment = 0.281 in; $a = 0.500$ in; $b = 0.375$ in; maximum load on spring = spring adjustment × rate; load S at nC diaphragm by equation 12, then with use of equation 11 P_1 is computed. Total differential of switch = sensitivity + differential adjustment. The differential spring rate at S from equation 15, $S/x = 123$ lb/in is shown in Fig. 7.

Toggle spring calculation where the geometry of the system is shown in Fig. 12.

k = spring rate

$$c = \sqrt{a^2+f^2-2af\cos\theta} \qquad (16)$$

$$f/\sin\gamma = c/\sin\theta \qquad (17)$$

The deflection of the spring dc, and the distance of travel at S, is x. The following equation results

$$S = kdc\,\sin\gamma\cos(\beta-\phi+\theta)a/b \qquad (18)$$

From equation 16

$$c = (a^2+f^2-2af\cos\theta)^{1/2}$$

$$dc = 1/2(a^2+f^2-2af\cos\theta)^{-1/2}(2af\sin\theta)d\theta$$

$$= \frac{af\sin\theta}{(a^2+f^2-2af\cos\theta)^{1/2}}d\theta \qquad (19)$$

also $x = b\sin(\beta-\phi+\theta)$

$$dx = b\cos(\beta-\phi+\theta)d\theta$$

$$\text{or } \cos(\beta-\phi+\theta) = dx/d\theta\,\frac{1}{b} \qquad (20)$$

From equation 17

$$\sin\gamma = f/c\,\sin\theta = \frac{f\sin\theta}{(a^2+f^2-2af\cos\theta)^{1/2}} \qquad (21)$$

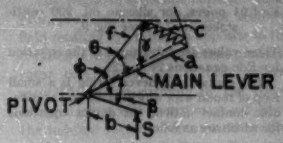


Fig. 12. Toggle spring calculation

Substituting equations 19, 20, and 21 in equation 18, the following results

$$S/dx = \frac{k(af/b\,\sin\theta)^2}{a^2+f^2-2af\cos\theta} \qquad (22)$$

where S/dx is the spring rate at S for any given value of θ. The value of the constants of the toggle spring system are: $a = 0.843$ in; $f = 0.265$ in; $b = 0.390$ in; $\alpha = 46.5$ degrees; $(\alpha-\theta) = 55$ degrees; and $\theta = 41$ degrees, initially. The toggle spring rate = 62 lb/in. These values are put in equation 22, and the resulting data are shown for S/dx in Fig. 7.

The capacity of the spring loaded nC diaphragm is shown as follows:

$$C = A_e^2/K \text{ in}^5/\text{lb/in}^2 \qquad (23)$$

The values of nC are as follows: $nC = 0.00452$, without differential spring load, for travel at $S = 0.03$ inch; and $nC = 0.00332$, with differential spring load, for travel at $S = 0.055$ inch. The nC characteristic is plotted in Fig. 7.

For the front C diaphragm spring refer to the foregoing as follows:

$nC = 0.00452$ and 0.00332 in⁵/lb/in²

let

$C = 0.01805$

so that

$n = 0.25$ and 0.18 respectively

From equation 23 $K = A_e^2/C = 83.5$ lb/in, and the spring rate = 81.1 lb/in.

Considering the resistance orifice R, and referring to Fig. 4, let $P_3/P_1 = 0.1$ for pump B, a representative pump. For $n = 0.18$, $\omega RC = 56$, and $\omega_p = 33$ rad/sec; see Appendix I. Then

$$R = 56/\omega_p C = 94 \text{ lb/in}^2\text{-sec/in}^3 \qquad (24)$$

Equation 3 is used

$$R = \frac{3.69\times10^{-2}\mu l}{r^4} \qquad (25)$$

For this orifice, then

$l = 0.172$ inch, length of orifice

$r = 0.019$ inch, radius for no. 62 drill hole

Table II. Calculation of Pump Discharge Pressure

Pump	C_h	R_h	L_h	Pounds per Square Inch P_C	P_R	P_L
A	14.2	0.00037	0.00405	0.0415	0.0095	5.37
B	60.7	0.00037	0.0044	0.0088	0.0066	2.52

$\eta = 1.94$ poises, for silicone oil, 200 centistokes at 77 F; see Fig. 9.

These values are substituted in the equation for R in equations 24 and 25.

The contact cantilever spring is a small loop of round wire which backs up the movable contact action, the design equations for which are as follows

$$L = \frac{11800d^3}{l} \text{ lb, maximum load}$$

$$g = \frac{0.00276l^3}{d} \text{ inch, deflection}$$

References

1. Controlling Pulsations in Hydraulic Equipment. C. E. Grosser. *Machine Design*, Cleveland, Ohio, vol. 14, Dec. 1942, pp. 108–12.

2. Communication Engineering (book). W. L. Everitt. McGraw-Hill Book Company, Inc., New York, N. Y., 1937, pp. 293–94.

3. Notes on Toggle Joints. *World Power*, London, England, Apr. 1935, pp. 174–78.

4. Electrical Contacts (book). G. Windred. D. Van Nostrand Company, Inc., New York, N. Y., 1940, p. 124.

5. Principles of Industrial Process Control (book). D. P. Eckman. John Wiley & Sons, Inc., New York, N. Y., 1945, pp. 103–19.

6. Dow Corning Silicone Notebook, Fluid Series No. 3, Midland, Mich., Sept. 1948, pp. 4, 5, 15. A bibliography on silicone publications is enclosed in this reference.

Power Supply for Commercial-Frequency Railroad Electrification

J. C. PRICE
MEMBER AIEE

COMMERCIAL-frequency railroad electrification, as a result of its increasing application in Europe, Africa, and elsewhere, has established its place as one of the prominent systems of electrification.

In planning commercial-frequency electrification, a major consideration is that of providing single-phase energy in large blocks to the motive power of the railroad. This energy must be provided economically from commercial 3-phase power systems without undue disturbance to the phase balance or wave form of the power system. It must not affect appreciably the performance of generation or utilization equipment connected to the power system.

This paper reviews the generic problems associated with provision of commercial-frequency power for traction purposes in single-phase blocks, and the technology and practices that have developed as an approach to these problems.

Nature of the Problem

Characteristics of Motive Power

The outstanding characteristic common to the types of motive power employed in the general case of commercial-frequency electrification is their use of single-phase power. While the provision of 3-phase energy to motive power has been studied, and is currently in use in Italy and Spain, practical experience indicates that the problems associated with provision of single-phase energy are less complex and less costly to solve than those associated with constructing and operating a 3-phase contact system for a main-line railroad electrification.

Through the efforts of organizations such as the International Electrotechnical Commission, certain features of commercial-frequency motive power are being standardized. For example, nominal values exist concerning the voltage of the contact systems which provide power to a-c electrifications. Three levels of rated voltage are considered as standard for single-phase networks: 11 kv, 15 kv, and 25 kv. For the purpose of this paper, discussion will be based upon a rated voltage of 25 kv, which is more likely to be used for future commercial-frequency electrification than other voltage levels.

Each of the several types of motive power employed on commercial-frequency electrifications has individual characteristics which influence the power-supply system. From a practical standpoint the types available for commercial application can be listed as those equipped with:

1. Series-wound a-c commutator motors.
2. D-c series-wound commutator motors fed by one or more.
   (*a*). Rectifiers.
   (*b*). Motor-generator sets.
3. Induction motors, fed by one or more of the following:
   (*a*). Phase changers.
   (*b*). Phase-frequency changers.

While induction-motored motive power has been employed in the past to a limited extent, the relative economics involved, primarily first costs, are such as to discourage any appreciable use of this type in the foreseeable future. Hence, further consideration of this type will be omitted from this discussion.

Either series-wound a-c motors or series-wound d-c motors fed by rectifiers are applicable to multiple-unit passenger cars as well as to locomotives. Because of space limitations under multiple-unit cars, the use of motor-generator equipment to feed d-c series-wound motors is restricted by practical considerations to locomotives.

Series-wound a-c motors present no untoward condition to the power supply system. Power factor of this type of motor ranges from 0.80 to 0.85 at loads above 50% (per cent) of its rated horsepower. Transformer tap changing is normally employed for speed control, although resistor control may also be used.

Rectifier motive power, with d-c traction motors, has a power factor characteristic in the range 0.80 to 0.85 at loads above 50% rated horsepower when speed control is obtained by transformer tap changing.

Fundamentally concomitant to the functioning of any rectifier, mercury-arc, semiconductor, or other type, with its associated load, is the harmonic current taken from the power system feeding the rectifier equipment. The magnitude of the harmonic current is an inverse function of the number of phases over which the rectifier load is divided, and obviously, the single-phase condition presents the potentially worst situation with respect to harmonic distortion.

Motive power employing a synchronous a-c motor driving a d-c generator presents an excellent load characteristic to the

Paper 56-977, recommended by the AIEE Land Transportation Committee and approved by the AIEE Committee on Technical Operations for presentation at the AIEE Fall General Meeting, Chicago, Ill., October 1–5, 1956. Manuscript submitted July 3, 1956; made available for printing August 3, 1956.

J. C. Price is with the General Electric Company, Schenectady, N. Y.

The author wishes to express his appreciation to Dr. J. E. Gerngross and T. F. Perkinson of the General Electric Company; to L. E. Jensen, formerly of the General Electric Company; and to others of his associates for their assistance in providing information which is incorporated in this paper.

Table I. Effect of Lower-Order Harmonics From Single-Phase Rectifier Locomotives

Order of Harmonic From Locomotive	Theoretical Magnitude*	Practical Magnitude†	Order of Harmonics Produced in Generator Rotor‡
Fundamental	1.000	1.000	2nd
3rd	0.333	0.20–0.28	2nd, 4th
5th	0.2000	0.13–0.16	4th, 6th
7th	0.143	0.05–0.10	6th, 8th
9th	0.111	0.01–0.07	8th, 10th

* Based on theoretical "flat-top" rectifier d-c output with no overlap between anodes.
† Determined from tests described by Hibbard, Whittaker, and Ames.[4]
‡ The order of harmonic produced in the generator rotor is influenced by transformer connections. Refer to Tables II and III.

power-supply system, in that there is virtually no wave-form distortion if the synchronous motor is properly designed, and adjustable excitation of the motor field permits maintenance of unity, or even leading, power factor.

The combination of d-c traction motors and a-c/d-c motor-generator set permits relatively easy procurement of regenerative braking with energy fed back to the contact system and through it to the power-supply system.

Effect of Single-Phase Loading on Generation Equipment

Single-phase loading of the power-supply system is a problem which has always been associated with a-c electrification, regardless of the frequency of the power system. Its solution has influenced the design of power-system equipment. A-c electrifications operating at 16⅔ cycles or 25 cycles, frequencies which are not at present widely used in applications other than traction, have generally been supplied with energy from special generators rated to supply single-phase loads. These generators, whether of the steam-turbine-driven type or of the type used in frequency-changer sets, have a single-phase rating which is appreciably below the capability of a machine of equivalent size handling a balanced 3-phase load.

The same factors which require 25-cycle generating equipment to be operated at less than its 3-phase rating when delivering single-phase power are also present in the case of commercial-frequency electrification.

The effect of single-phase loading upon the capability of a power system and its associated equipment may be visualized by applying the method of symmetrical components to a situation involving a single-phase traction load.

A single-phase load (or load unbalance) of, for example, 10,000 kva may be resolved by the method of symmetrical components into two 3-phase components. These components, as shown in the appendix, correspond to the single-phase kva, and in this case are

1. A system of positive-sequence currents corresponding to 10,000 kva.
2. A system of negative-sequence currents corresponding to 10,000 kva.

The positive-sequence currents appear like a balanced 3-phase load of 10,000 kva to the alternator, and produce no unusual effects.

The negative-sequence currents, however, produce in the alternator stator a field rotating in a direction opposite to the rotation of the rotor. This field produces rotor core losses and it also induces second-harmonic currents in the rotor. In a generator with amortisseur windings, the induced rotor currents cause heating of the amortisseur windings. In a non-salient-pole generator which does not have amortisseur windings, the induced currents cause heating of the rotor surface,[1] particularly in the vicinity of slot wedges at the ends of the rotor.

Relatively little information is available today concerning permissible negative-sequence loading of generating equipment that does not have amortisseur windings. Modern high-speed alternators of the non-salient-pole type are not designed specifically to handle negative-sequence loading of any appreciable magnitude on a nontransient or continuous basis.

Effect of Single-Phase Loading on Utilization Equipment

With commercial-frequency electrification, the effect of single-phase loading upon the utilization equipment of other users of electric energy becomes an important consideration. It is evident that this was a considerably less important factor with special-frequency electrifications where few, if any, commercial users of power were fed from the same power system as the railroad. With the application of commercial-frequency power to single-phase traction contact systems, however, it becomes necessary to maintain voltage balance at the delivery to other power users within limits which their equipment can tolerate.

Therefore, voltage unbalance present at points on the power system must

Table II. Rectifier Locomotive Harmonics: Effect of Balancing Connections Upon Harmonic Behavior

Order of Harmonics	Two Equal Loads Connected to 3-Phase System Through a Scott Transformer Connection		Three Equal Loads Each Connected to a Separate Phase of the 3-Phase System		Six Equal Loads Connected to 3-Phase System Through Three Scott-Connected Transformers, Each Displaced 120 Degrees From the Other Two	
	Sequence of Harmonic Current	Effect of Harmonic Current	Sequence of Harmonic Current	Effect of Harmonic Current	Sequence of Harmonic Current	Effect of Harmonic Current
Fundamental	Positive		Positive		Positive	
3rd	Negative	Produces 4th harmonic in generator rotor	Zero	May be eliminated by circuit connections*		Cancelled by circuit connection
5th	Positive	Produces 4th harmonic in generator rotor	Negative	Produces 6th harmonic in generator rotor		
7th	Negative	Produces 8th harmonic in generator rotor	Positive	Produces 6th harmonic in generator rotor		
9th	Positive	Produces 8th harmonic in generator rotor	Positive	May be eliminated by circuit connections*		
11th	Negative	Produces 12th harmonic in generator rotor	Negative	Produces 12th harmonic in generator rotor	Negative	Produces 12th harmonic in generator rotor

* The use of one or more wye-connected transformer windings, without neutral connection, or the use of one or more delta-connected windings both of which present an open circuit to zero-sequence current, will eliminate zero-sequence third-order currents in the alternator.

This case is theoretical and assumes that the generator is influenced equally by the load at each step-down transformer.

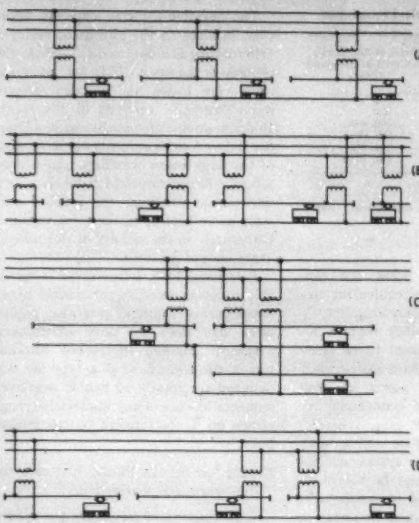


Fig. 1. Methods of distributing single-phase electrification load among the three phases of a commercial power system

A—Center feed
B—End feed, two points
C—End feed, one point
D—End feed, one point

be considered when applying a load with a single-phase component to the system. The approximate voltage unbalance at a point on the power system may be determined from the relation:

$$\text{Per cent voltage unbalance} = \frac{\text{negative-sequence kva} \times 100}{\text{available short-circuit kva}}$$

The effect of voltage unbalance upon a specific motor can be determined by its designer and a relationship established between loading of the motor and permissible voltage unbalance, other factors being assumed normal. Some of the factors involved in determining the ability of a polyphase motor to operate under unbalanced voltage conditions are illustrated in the following discussion of the effect of unbalance upon polyphase induction motors.

When the line voltages applied to a polyphase induction motor are not equal, unbalance currents will flow in the stator winding, the magnitude of such current depending upon the degree of unbalance. A small amount of voltage may increase the current an excessive amount.

The effect of unbalanced voltages on polyphase induction motors is equivalent to the introduction of a negative-sequence voltage having a rotation opposite to that occurring with balanced voltages. This negative-sequence voltage produces in the motor air gap a flux rotating against the rotation of the rotor, tending to produce high currents similar to those involved in "plugging" the motor. A small negative-sequence voltage may produce in the stator windings currents considerably in excess of those present under balanced voltage conditions.

The voltage unbalance in per cent is defined as follows:

$$\text{Per cent voltage unbalance} = \frac{100 \times \text{maximum voltage deviation from average voltage}}{\text{average voltage}}$$

For example, with separate phase voltages of 210, 215, and 220, the average would be 215 volts, while maximum deviation from the average would be 5 volts. The unbalance would be

$$\text{Per cent voltage unbalance} = 100 \times \frac{5}{215} = 2.32\%$$

A relatively small unbalance in voltage will cause a considerable increase in temperature rise. In the phase carrying the greatest current, the percentage increase in temperature rise will be approximately two times the square of the percentage voltage unbalance. The increase in losses and the consequent increase in average heating of the entire winding will be slightly lower than in the winding with the greatest current. For example, a 3½% unbalance will cause approximately 25% increase in temperature rise in the phase carrying the largest current. That is,

$$\text{Per cent increase in temperature rise} = (3\tfrac{1}{2})^2 \times 2 = 24.5\%$$

For a 5% unbalance,

$$\text{Per cent increase in temperature rise} = (5\%)^2 \times 2 = 50\%$$

The locked-rotor torque, the breakdown torque, and the full-load speed are reduced when the induction motor operates on unbalanced voltages.

The locked-rotor stator currents will be unbalanced to the same degree that the voltages are unbalanced, but the kilovolt-amperes under these conditions will be increased only slightly.

Currents at normal operating speed with unbalanced voltages will be greatly unbalanced, in the order of six to ten times the voltage unbalance. This introduces a problem in selecting the proper overload protective devices.

In the case of synchronous motors, it is reasonable to expect that an investigation by the designer to determine the effect of unbalanced supply voltages would indicate ability of the motor to tolerate a small voltage unbalance while operating under rated conditions. However, this is not a general guide, and each motor must be considered individually to arrive at a definite answer.

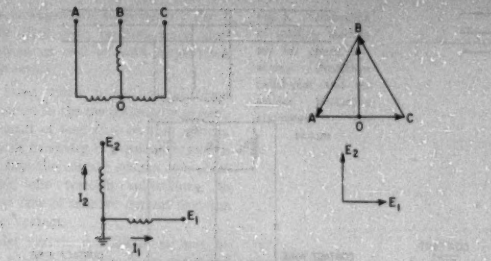
The foregoing discussion indicates that customer sensitivity to voltage unbalance is a factor which electrical utility companies supplying a commercial-frequency electrification must consider.

Effect of Harmonic Currents on Generation and Utilization Equipment

The trend which is evident in the United States and in Europe toward increasing use of rectifier-type motive power, coupled with the rapid pace of rectifier development, indicates that future electrification systems will have to be planned to handle rectifier-type motive power.

A fundamental characteristic of rectifier circuits is that they generate harmonic currents in the a-c power system from which they are energized. This is a well-known phenomenon and it has been exhaustively treated in technical literature. For the purpose of this paper it is proposed to limit the discussion to the harmonic values associated with rectifiers energized from a single-phase source, and operating with no phase retard.

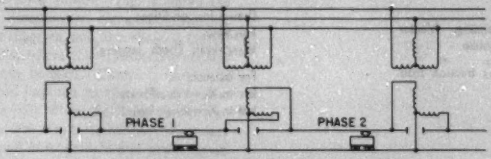
Table I indicates the magnitude of lower-order harmonic currents taken by



(A) SCOTT TRANSFORMER CONNECTION AND VOLTAGE DIAGRAM

BALANCED 3PH LOADING RESULTS WHEN $I_2 = jI_1$

3PH PF = 2PH PF UNDER BALANCED CONDITIONS



(B) APPLICATION OF SCOTT CONNECTED TRANSFORMERS TO ELECTRIFIED RAILWAY

Fig. 2. Scott-connected transformers for distributing single-phase traction load among three phases of commercial power systems

single-phase rectifier locomotives. The theoretical magnitude is obtained when the d-c circuit has infinite inductance and the a-c circuit has zero inductance and no losses, under which condition each harmonic has a per-unit value equal to the reciprocal of the order of the harmonic. This is the condition of "flat-top wave with no overlap" to which reference is made in the table.

Stator losses in conventional rotating machinery such as generators, synchronous motors or condensers, and induction motors, are increased by the presence of harmonic currents. The increased losses include increased I^2R losses in stator copper, and increased hysteresis and eddy current losses in the stator magnetic circuit. Because hysteresis loss is a linear function of frequency and eddy current loss is a square function of frequency, the effect of a given amount of harmonic current upon stator core losses increases exponentially as the order of the harmonic becomes higher.

Harmonics also cause increased rotor losses in rotating machinery. It is evident that a rotating field of harmonic frequency in the stator of a machine will induce current in the rotor circuit. In the case of rectifier load, the harmonics in the stator circuit will be odd-order harmonics, and the currents induced in the rotor circuit will be even-order harmonics. The rotor harmonics will be of the next-higher order if the stator harmonic current is negative-sequence, and of the next-lower order if it is positive-sequence.

Table II indicates the order of harmonic which may be produced in the rotor circuit by each of the lower-order rectifier harmonics. Specific conditions which influence the order and the magnitude of harmonic current produced in the rotor will be discussed under "Effect of Phase Balancing Upon Harmonics."

Core losses and circulating current losses result in the rotor from the presence of harmonic flux in the rotor. The effects are similar to those produced by the double-frequency field in the rotor associated with negative-sequence current of fundamental frequency in the stator. The relative effect increases with increasing order of the harmonic, because core losses increase exponentially with frequency and because the X/R ratio of amortisseur windings or squirrel-cage windings increase and concentrate current nearer the surface of such windings as frequency increases.

Harmonic currents may influence the power system and may influence associated equipment other than that mentioned. However, the effects upon rotating machinery appear to be of potentially greater significance in the economics of commercial-frequency electrification, and discussion was for that reason limited to the effects upon motors and alternators.

Approaches to the Problem

To Balance or Not to Balance?

The answer to this question may have appreciable influence on the economic attractiveness of a commercial-frequency electrification. For this reason it is highly important that all of the factors influencing the need to balance be accurately known and carefully evaluated. The effect of unbalanced conditions is more critical in rotating equipment than it is in other power-system elements, and this discussion will be limited to considering this question in terms of requirements imposed by generators, motors, and synchronous condensers.

At this point it is necessary to inquire concerning how balancing may be accomplished if it proves to be necessary.

The general approach to balancing the single-phase traction load has two principal aspects: (1) using the inherent balance between loads on different sections of the railroad to secure partial balance of the load presented to the 3-phase system; and (2) if necessary, further refining the 3-phase balance by means of special balancing equipment connected to the traction power system.

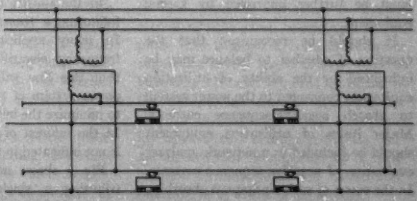


Fig. 3. Scott-connected transformers for 2-track commuter electrification

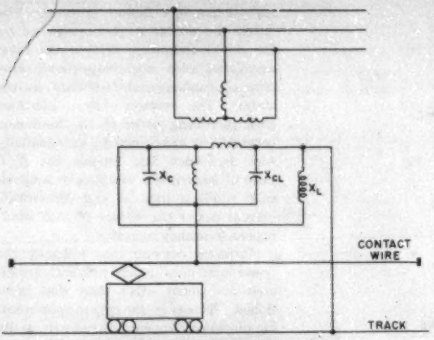


Fig. 4. Scott-connected transformers for static balancing. Vector sum of 2-phase voltages applied to contact system

X_{CL} = impedance for power-factor correction of lagging traction load
X_C X_L, = balancing kva elements

For balance:
kva in $(X_C+X_L+X_{CL})$ = load kw + load kvar

For 3-phase unity power factor:
kva in X_C = kva in X_L = 1/2 load kw
kva in X_{CL} = load kvar

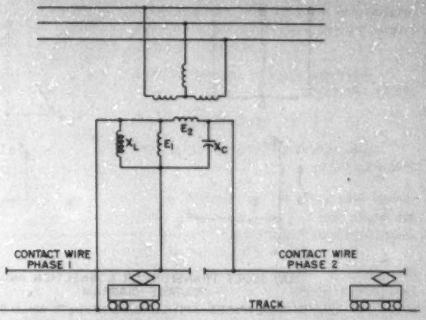


Fig. 5. Scott-connected transformer for static balancing. Each of 2-phase voltages applied to separate section of contact system

Drawn for condition:
$kw_2 > kw_1$
$kvar_2 > kvar_1$ (both lagging)

For balance:
kva in X_L = $kvar_2 - kvar_1$
kva in X_C = $kw_2 - kw_1$

Considering the magnitude of traffic density which is required to justify electrification today, it is practically certain that the first of these approaches will be used in all cases except possibly those of an experimental nature. The question of whether to balance thus becomes one of whether special equipment is required to refine further the inherent balance. This can be resolved by a study to determine the effect upon the power supply system of specific railroad load conditions and the unbalance which they produce. A network analyzer study will reveal the negative-sequence loading of generating equipment and the voltage unbalance at critical points on the power system. A comparison of these quantities with the capabilities of specific pieces of equipment will reveal to what extent the inherent balance of the system must be further improved by special equipment.

It should be recognized that the question of whether to balance may be influenced by the ability of utilization equipment connected to the power system to absorb negative-sequence currents. Major items of utilization equipment should be included in a network analyzer evaluation of negative-sequence currents. In the event of such equipment absorbing negative-sequence current without encountering the limitations described under "Effect of Single-Phase Loading on Utilization Equipment," it will be a useful addition to the negative-sequence "sink."

With this introduction to the question as background, further attention will now be devoted to methods by which phase balancing may be accomplished.

Utilizing Inherent Balance

This method is relatively economical because it takes advantage of balance already inherent in the railroad's operation. Essentially it consists of segregating the traction distribution system into sections and distributing power feed to these sections among the several phases of the power-supply system. How this may be accomplished will now be examined in greater detail.

In the illustrations referred to in the following, a transmission line is shown. In many applications of commercial-frequency electrification it is expected that this line will be replaced by the power system of the electricity supplier, to minimize the initial cost of electrifying. In the interest of simplicity, switchgear is not indicated in these illustrations.

Fig. 1 shows several methods of distributing a single-phase electrification load among the three phases of a commercial-frequency system.

Fig. 1(A) shows how an electrification may be energized from a number of single-phase substations, with one third of the substations being energized from each phase of the supply system. With this method of feeding the line, isolation of phases is provided at some point between the substations.

In Fig. 1(B), each section of the contact system is fed in multiple from adjacent substations. In this case, phase isolation is provided at each substation. Where a multiple-feed system is employed, the section of the contact system so energized must be connected to sources of the same phase position, in order to avoid heavy circulating currents.

The arrangement of Fig. 1(C) would find application in a main-line electrification where two branches are electrified, or in a commuter operation where several lines radiate from a single point.

There are many possible combinations of various simple methods of feeding the line, and Fig. 1(D) shows one combination which has found application in Europe. In this case, the overhead contact system is divided into three sections, and one section is energized from a single-transformer substation at one extremity of the electrification, while the other two sections are energized from a 2-transformer substation located about one third of the distance from the other extremity of the electrification.

The arrangements suggested by Fig. 1 can be used with multiple-track electrification, as well as with single-track electrification.

In cases where a substantial phase shift occurs in the electric supply system, as a result of load flow on that system, it may be necessary to introduce a section break into the contact system at a point between the traction substations, to prevent flow of reactive current over the contact system.

Under certain conditions, it may be more practical to use Scott-connected transformers to provide distribution of the traction load among the three phases of the commercial-frequency supply, than to accomplish this by energizing various sections of the contact system from each of the three phases. Fig. 2 shows the Scott transformer connection and its vector diagram, and also shows a typical situation where this transformer connection may be used. Complete balance of the load imposed on the 3-phase system results only when the loads on the two legs of the Scott transformer secondary are equal in magnitude and power factor. Section breaks may be required between substations if large phase shifts exist in the primary supply system.

In Fig. 2, if the load consisted of a relatively large number of motive-power units, and the loading in each section were approximately equal, the Scott transformer connection would be advantageous, and would reflect only a relatively small single-phase unbalance to the 3-phase system. The Scott transformer connection may also be applied on a relatively short electrification where it is practical to feed the contact system at a single point near the center of the electrification. With relatively dense traffic, the traction loads on the two sections of the contact system could be nearly equal, and a limited amount of single-phase unbalance would be reflected into the 3-phase system.

Another variation of the Scott transformer connection is shown in Fig. 3. In this arrangement each track in a 2-track railroad is energized from a separate phase of the 2-phase output of the transformer. This type of balancing might find application in a high-density commuter operation where there would be inherently a reasonable degree of balance between the two secondary single-phase loads and where the power system in a metropolitan area would be able to tolerate a certain degree of unbalance.

The simple Scott transformer connection as shown in Figs. 2 and 3 would probably not be effective in distributing load among the three phases of the commercial system if the load consisted of only a small number of motive-power units. It appears best adapted to a suburban type of operation where a large number of motive-power units are in operation on the system, or for a main-line electrification with frequent freight- and passenger-train service.

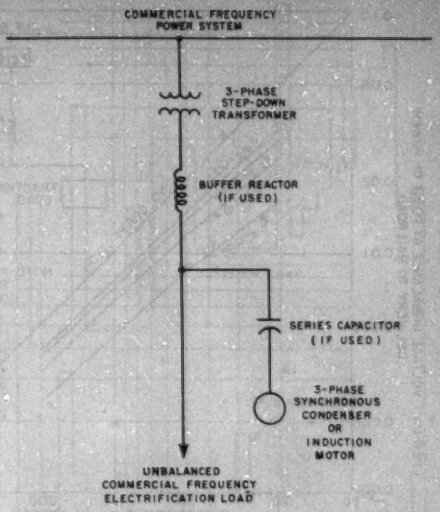

COMMERCIAL FREQUENCY POWER SYSTEM
3-PHASE STEP-DOWN TRANSFORMER
BUFFER REACTOR (IF USED)
SERIES CAPACITOR (IF USED)
3-PHASE SYNCHRONOUS CONDENSER OR INDUCTION MOTOR
UNBALANCED COMMERCIAL FREQUENCY ELECTRIFICATION LOAD

Fig. 6. Application of rotating machinery for phase balancing. Protective switchgear and capacitor protective equipment not shown

Balancing Methods

Where local conditions require a greater degree of 3-phase balance than can be provided by utilizing the inherent balance in the railroad operation, it becomes necessary to remove or reduce the "residual" phase unbalance.

The method and point of application of balancing provisions depend upon the equipment which must be protected. The provision of amortisseur windings in generating equipment would appear to be an ideal solution where considerations of voltage unbalance at utilization equipment are not limiting. The design and cost of such amortisseur windings will be affected by the degree of unbalance specified. This solution, it is evident, would result only from long-range planning undertaken jointly by the railroad to be electrified and by the supplier of electric energy.

Another possible solution where utilization equipment is not limiting, and where generators do not have amortisseurs, is to locate balancing equipment at critical points on the power system to protect generating equipment.

The third approach to correcting a residual unbalance condition is the application of balancing equipment at the point of power delivery to the railroad substation or transmission circuit. Both rotating equipment and static equipment have been considered for this service.

Circuits Employing Static Equipment

The use of static equipment has been suggested for phase balancing because of the possibility of lower first cost than rotating equipment.

Static balance circuits employ a Scott transformer connection to provide a circuit which is capable of completely balancing the load imposed upon the 3-phase system when proper voltage and current relations exist in the 2-phase secondary circuit. Static capacitors and reactors are used to provide the voltage and current conditions which are necessary in order for the Scott transformer to provide complete balance. A fundamental relation when balancing a single-phase load through a Scott transformer connection is that the total kva of active balancing element required is equal to the arithmetical sum of the load kw plus the load kilovars (kvar). Depending upon circuit arrangements, the balancing kva can be supplied by a combination of inductance and capacitance elements, or by either of these alone.

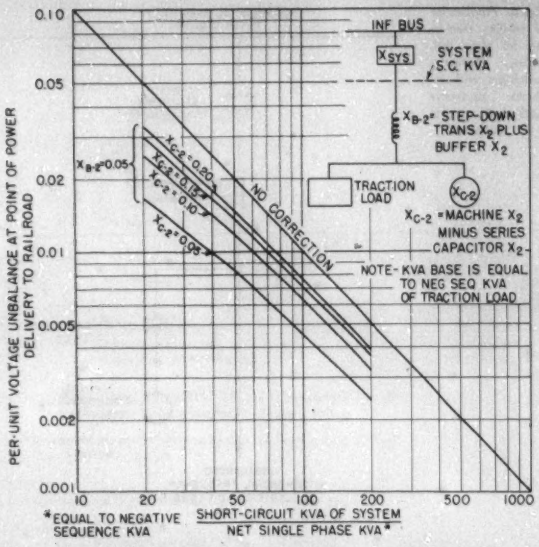


Fig. 7. Effect of rotating balancing equipment upon voltage unbalance

Fig. 4 shows a Scott-transformer balancing circuit where the voltage applied to the single-phase load is the vector sum of the individual secondary voltages. This balancing circuit appears to be best suited for main-line electrifications where there is little diversity of traffic to provide any inherent balance on the 3-phase system.

In this circuit, the power factor of the single-phase load must be corrected to unity before 3-phase balance can be achieved. At unity power factor, balanced loading on the 3-phase side is obtained when

Capacitor kva = reactor kva = 1/2 load kw

Where some degree of balance exists between two line sections because of diversity of traffic and relatively large numbers of trains in service, a connection of static balancing elements can be used with the Scott-transformer connection as shown in Fig. 5. This circuit will be recognized as the Scott-connected transformer of Fig. 2 or Fig. 3 with balancing elements added. In this arrangement, only that part of the traction load which is not balanced by the physical arrangement of the contact system need be balanced by the static elements. Again, the fundamental relation that balancing kva equals unbalance kw plus unbalance kvar applies.

Application of equipment for static phase balancing remains to be worked out in detail. Some of the problems involved include:

1. Determining the most practical methods for continuously switching or varying the reactive kva provided by large banks of capacitors and reactors.

2. Determining the degree of unbalance which the 3-phase commercial systems can tolerate, both on a continuous basis and on a short-time or momentary basis while the balancing circuit is taking corrective action.

3. Designing a control system to include both power-factor correction and phase balancing, to provide the required accuracy and speed of response.

4. Determining the effect of harmonic currents, if present in the traction load.

Phase-Balancing Circuits Employing Rotating Equipment

The use of rotating machines as phase converters to balance a single-phase load was discussed in technical literature as early as 1916,[3] and probably before that. Recent investigations, however, have been concerned with evaluating the use of machines of relatively standard types, synchronous condensers and squirrel-cage induction motors, rather than the relatively complex machines that were envisioned and constructed in connection with electrification activity of a generation or more ago.

An approach to the phase-balancing problem employing rotating machinery, either synchronous condensers or induction motors, has been considered and may find economic application where it is desired to remove negative-sequence currents at the point of power delivery to a traction stepdown substation or to a 3-phase transmission line which energizes several railroad substations. This arrangement is shown in Fig. 6.

The rotating machine provides a negative-sequence sink, capable of absorbing negative-sequence kva from the traction load. The buffer reactor influences the proportion of negative-sequence current flowing to the power system. The series capacitor, if used, increases the negative-sequence loading of the rotating machine by decreasing the effective impedance of the sink.

Table III. Rectifier Locomotive Harmonics: Action of Harmonic Content of Unbalanced Load

	One Phase Only of 3-Phase Power System Loaded		One Phase Only of 2-Phase System Loaded (Where Scott-Connected Transformers Are Used)	
Order of Harmonic	Sequence of Harmonic Current	Effect of Harmonic Current	Sequence of Harmonic Current	Effect of Harmonic Current
Fundamental	All single-phase currents	Each of these single-phase currents may be resolved into a positive-sequence current and a negative-sequence current	All single-phase currents	Each of these single-phase currents may be resolved into a positive-sequence current and a negative-sequence current
3rd 5th 7th 9th 11th etc		Positive-sequence currents produce the $(n-1)$ harmonic, and negative-sequence currents the $(n+1)$ harmonic in the rotor, where n is the order of the harmonic in the stator		Positive-sequence currents produce the $(n-1)$ harmonic, and negative-sequence currents the $(n+1)$ harmonic in the rotor, where n is the order of the harmonic in the stator

Fig. 7 indicates the improvement in voltage unbalance which may be realized with various combinations of buffer reactor and series capacitor, as a function of

$$\frac{\text{System short-circuit kva}}{\text{Negative-sequence kva of load}}$$

Because the negative-sequence reactance of a synchronous condenser or of a large induction motor is very nearly equal to its subtransient reactance, this system is sensitive to transient conditions, and would require capacitor protective equipment as is normally used in series-capacitor applications.

The series-capacitor–synchronous-condenser arrangement is maximum-rated for practical purposes, because of the limited thermal capacity of the amortisseur winding which will have induced in it second-order harmonics. In a recent investigation for 50-cycle application, it was estimated that it would be necessary to derate the synchronous condenser to about 50% of its equivalent 3-phase rating and to provide it with a special, heavy amortisseur winding. At 25 cycles, the derating would be to approximately 60% of the machine's equivalent 3-phase rating, and at 60 cycles to about 40% of the same rating.

As in the case of static balancing equipment, it is necessary to determine the solution for several problems apparent in this balancing arrangement if the application of capacitors is to be considered. These include:

1. Protection of capacitor against overvoltages resulting from surges.
2. Providing positive damping for low-frequency currents.
3. Possibility of ferroresonance when energizing a synchronous condenser through a series capacitor.
4. Possibility of hunting of the machine.
5. Degree of compensation which is practicable.

These and other problems are discussed in recent technical literature.[8]

Effect of Phase Balancing upon Harmonics

Consideration has been given to two methods of utilizing the inherent balance of traffic between different sections of a railroad to obtain a degree of balance of loading among phases in a 3-phase power system:

1. Energizing sections of the electrified operation from different phases of the 3-phase power system.
2. Energizing sections of the electrified operation from different phases of a 2-phase power supply, which in turn is energized from the 3-phase power system through a Scott-connected transformer.

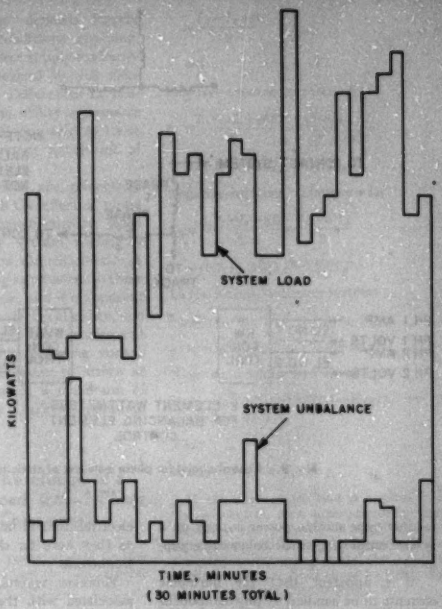


Fig. 8. System load and system unbalance for a commuter operation. Load data shown for first 30 minutes of 2-hour afternoon peak load period

In either case, it is recognized that exact balance will result at best on a momentary or relatively short-duration basis. For practical purposes, the usual situation with either of the above methods will be:

1. Part of the load reflected in the 3-phase system will be balanced, and may be represented as a normal positive-sequence load.
2. The remainder of the load will be unbalanced, and may be considered as a single-phase load, which, in turn, may be represented as a system of equal positive-sequence currents and negative-sequence currents.

Table II shows the effect of the two balancing connections upon the harmonics associated with the balanced portion of the load. Comparison of the first two columns of this table indicates that direct distribution of load among the three phases of the a-c power system may permit eliminating third-order harmonics from the generating equipment and utilization equipment, while a single Scott-connected transformer does not offer this possibility. However, it must be recognized that limitations exist in the former method, because of the impracticability in many cases of loading the three phases of the power system at one point; this limits the maximum reduction in third-order harmonics to those parts of the power system which are influenced by the loading at several substations.

The third column of Table II shows that additional harmonics may be eliminated by connecting six equal loads to the power system through three Scott-connected transformers, each of which is displaced 120 degrees electrically from the other two.

Table III shows the effect of the unbalanced portion of the traction load.

Phase-balancing equipment, if it is required in a specific application, may affect the amplitude of harmonic currents. For example, rotating phase-balance equipment on the 3-phase system would be expected to absorb harmonic current in some degree.

Balancing equipment consisting of static capacitors and reactors would be expected to influence harmonic conditions but in a manner which is not predictable as a general case. Because of the possibility of resonance occurring at harmonic frequencies in such a circuit, the use of harmonic filters may be necessary when

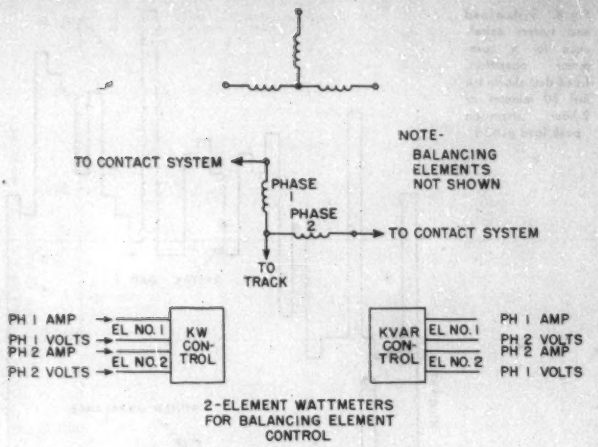


Fig. 9. Control circuit for phase balancer of static type

rectifier-type motive power is used on a system employing static balancing equipment.

It is apparent that the harmonic currents to be handled by any component part of a power system are influenced not only by the amplitude of harmonic currents generated by a unit of motive power, but also by the degree of balance in the traction power system and by the method of connecting that power system to the power system of the electric-energy supplier. These are prime considerations; other factors such as the nature of balancing equipment, and the possible use of harmonic filters, will also be influencing factors.

Application Engineering of the Power Distribution System

To an important degree, the application engineering problems encountered in planning the distribution system for a commercial-frequency railroad electrification are similar to those encountered in planning an a-c electrification of lower frequency and lower voltage.

It is not within the scope of this paper to discuss the many considerations involved, for example, in determining whether return current shall flow entirely through the rail or whether a 3-wire system shall be used; or in determining how the feeder breakers shall be relayed; or in specifying a feeder breaker. These questions, and similar questions, may be resolved for commercial-frequency electrification in much the same manner as they were for electrification at lower frequencies.

However, several special problems are associated with the distribution system for commercial-frequency electrification, and they will be discussed in the following paragraphs.

Pattern of the Loading of Phase-Balance Equipment

Where the arrangement of traction power supply is such that balancing equipment must handle only a residual unbalance, the pattern of the load imposed upon balancing equipment may be greatly different than the traction load pattern.

This is illustrated by Fig. 8 which shows system load and system unbalance for a commuter operation. These data are based upon a proposed system to be energized from a Scott-connected transformer, and with one phase feeding each of two tracks. Transposition of phases was proposed to minimize unbalance resulting from fleet operation of trains at the start of the morning and evening rush periods.

In cases such as this the loading of balancing elements may be characterized by high peaks of relatively short duration (as compared with the duration of peaks in the traction load), and this will influence the ratings required for balancing-circuit components such as transformers, reactors, or other devices which have significant thermal-storage capacity.

Ability of Balancing Elements to Handle Peaks of Short Duration

The prime function of balancing elements in phase-balancing circuits is to provide energy storage for a fraction of a cycle to convert the pulsating energy flow characteristic of single-phase power to the uniform energy flow of 3-phase power. Despite their ability to provide energy storage, however, certain balancing elements, shunt capacitors and the amortisseur windings of rotating machinery in particular, have relatively little ability to handle overloads of short duration. In general, the permissible overload-versus-time characteristic for either of these elements will not permit significant overloading in phase-balancing service. This means from a practical viewpoint that the continuous capacity of these phase-balancing elements must be adequate to handle the peak-balancing requirements, even if they are of short duration.

Economical Voltage Level for Operation of Balancing Equipment

It is reasonable to expect that the majority of commercial-frequency electrifications both in the United States and abroad will have a contact system voltage of 25 kv and will receive power from electrical utility systems at voltage levels appreciably higher than this. If electrification were to develop today in the United States, for example, it would be practical in many areas to make delivery of power to traction substations at a voltage of 132 kv.

On the other hand, practical phase-balancing equipment will generally have an operating voltage in the range from 4,160 to 13,800 volts. This indicates the need for step-down transformers or auto-transformers to provide the proper voltage level for phase-balancing equipment. Such transformers can be selected to take advantage of the load pattern characteristic of phase-balancing operation, discussed previously.

The particular voltage level to be used for a rotating phase balancer of a given kva rating will depend largely upon the relative values of over-all cost at each voltage level, considering the step-down transformer, switchgear, rotating machine and its starting equipment, and series capacitors.

In the case of static balancing-equipment, availability of adequate and economical equipment for adjusting balancing capacity may be the determining factor in establishing the voltage level. When estimates of a preliminary nature

are made, conventional switching devices which have been applied in utility or industrial capacitor switching applications are usually considered to establish order of magnitude phase-balancing costs. It should be emphasized that the need to make a thorough investigation of the performance of these devices in phase-balancing service has not yet occurred, and that such investigation is a prerequisite to application in this service.

Control of Balancing Equipment

Balancing circuits which employ rotating machinery are self-regulating with respect to balancing action. The negative-sequence kva absorbed by the machine is in proportion to the total negative-sequence kva generated by the load.

Static balancing circuits require regulating equipment to initiate balancing action. For the purpose of determining the approximate cost of control for the static balancing system of Fig. 5, a control circuit employing 2-element wattmeters was devised. This system is illustrated in Fig. 9. The kw and the kvar in each phase of the 2-phase systems are measured in separate 2-element wattmeters. These wattmeters are connected so that unbalance between phases in either of these quantitites will initiate corrective action in the proper balancing element. This action tends to create a balanced condition in the 2-phase system, which, in turn, will be presented as a balanced load to the 3-phase system by the Scott-connected transformer.

Conclusions

The power-supply problems associated with commercial-frequency railroad electrification are of a nature that requires each case to be resolved on its own merits. Whether the load of the electrified railway can be successfully integrated with the power-supply system depends largely upon the specific conditions applying. Considerations of power supply evidently are not limiting in several 50-cycle electrifications abroad. Differences in railroad operation and in utility systems in the United States as compared to those abroad, however, present a new set of conditions to investigate.

The uncharted areas are principally those concerned with the effect of phase unbalance and harmonic currents upon the electical utility system feeding a commercial-frequency electrification. A variety of engineering approaches to these problems is discernable, and it is apparent that the railroad, the electrical utility, and the manufacturers of electric equipment all have important engineering contributions to make in order to arrive at solutions which provide a maximum of incentive to electrify a railroad operation.

Appendix. Resolution of a Single-Phase Load into 3-Phase Sequence Components

Given: A 3-phase power system feeding a single-phase load so that the line amperes in the 3-phase system are

$$I_a = I_A + j0$$

$$I_b = -I_A + j0$$

$$I_c = 0$$

and the 3-phase line kv are

$$E_{ab} = E_{bc} = E_{ca} = E_{AB}$$

The kva of the single-phase load is

$$\text{kva}_{AB} = E_{AB} I_A$$

By the method of symmetrical components

Zero-sequence amperes $= I_{s0}$

$$I_{s0} = \frac{I_a + I_b + I_c}{3}$$

$$= \frac{I_A - I_A + 0}{3}$$

$$= 0$$

Positive-sequence amperes $= I_{s1}$

$$I_{s1} = \frac{I_a + aI_b + a^2 I_c}{3} = \frac{I_A - aI_A}{3}$$

$$= (0.5 - j.289) I_A = 0.577 I_A$$

Negative-sequence amperes $= I_{s2}$

$$I_{s2} = \frac{I_a + a^2 I_b + aI_c}{3} = \frac{I_A - a^2 I_A}{3}$$

$$= (0.5 + j.289) I_A = 0.577 I_A$$

In the 3-phase sequence systems

Positive-sequence kva $= \text{kva}_1$

$$\text{kva}_1 = \sqrt{3}\ E_{AB}(0.577 I_A) = E_{AB} I_A$$

Negative-sequence kva $= \text{kva}_2$

$$\text{kva}_2 = \sqrt{3}\ E_{AB}(0.577 I_A) = E_{AB} I_A$$, on an equivalent basis.

Thus

$$\text{kva}_{AB} = \text{kva}_1 = \text{kva}_2$$

If the single-phase load is applied from line to neutral, then a zero-sequence component is also present. It may be shown by analysis similar to the foregoing, that the positive-sequence kva and the negative-sequence kva are both equal to the single-phase kva in this case as well.

References

1. Operation of Nonsalient-Pole-Type Generators Supplying a Rectifier Load, M. D. Ross, J. W. Batchelor. *AIEE Transactions*, vol. 62, Nov. 1943, pp. 667–70.

2. Single-Phase Power Production, E. F. W. Alexanderson, G. H. Hill. *Ibid.*, vol. 35, pt. II, 1916, pp. 1315–27.

3. Technical Problems Associated with the Application of a Capacitor in Series with a Synchronous Condenser, R. L. Witzke, E. L. Michelson. *Ibid.*, vol. 70, pt. I, 1951, pp. 519–28.

4. Rectifier-Type Motive Power for Railroad Electrification, L. J. Hibbard, C. C. Whittaker, E. W. Ames. *Ibid.*, vol. 69, pt. I, 1950, pp. 519–23.

Discussion

L. B. Curtis (The Pennsylvania Railroad, Philadelphia, Pa.): Mr. Price has made a valuable contribution to railroad electrification progress with his excellent paper. This study is of immediate practical value in Europe and other countries in the world where commercial-frequency electrifications are already in operation, are being built, or are in the planning stage. In the United States the study is at present academic, but in the foreseeable future, due to the rising costs of oil and diesel-electric locomotive maintenance, railroad electrification of heavily traveled main track must again be considered; and the study will then be of great value. Of all the electrification systems now in existence, high-voltage commercial-frequency electrification appears to offer the most economical solution.

Mr. Price ably points out from a theoretical point of view the possible additional heating effect of phase unbalance, with its resultant negative-sequence currents, on generating apparatus as well as on utilization apparatus of other users of electric energy, due to imposing single-phase railroad loads on a commercial 3-phase system. It should be pointed out that the methods of distributing a single-phase electrification load among the three phases of a commercial-frequency system as illustrated in Fig. 1 assume a transmission system paralleling the railroad. This is a simplification that will not always obtain, as in many cases the railroad trolley system would be tied to transmission lines, at approximately 50-mile intervals, that either cross the railroad or come close to it. These lines may or may not be at present interconnected, and will certainly not always be part of the same ownership. The extra cost of keeping the 3-phase system within safe limits of balance may well affect the decision as to whether to electrify or not.

Committee 13, Railway Electrification, of the Electrical Section of the Association of American Railroads, has been studying all aspects of the commercial-frequency

type of electrification for over a year. One task group in particular is making a practical investigation of the effect of phase unbalance due to railroad single-phase loads on an existing system. A 200-mile section of the Norfolk and Western Railway between Williamson, W. Va., and Roanoke, Va., was selected for study. The train schedules were translated into electrical loads which were furnished to the American Gas and Electric Company for their determination of the effect of these single-phase loads on their system. The General Electric Company is making tests on a generator being built by them for this system to determine the tolerable limits of phase unbalance. While the studies and tests are not concluded as yet, preliminary results indicate that the resultant unbalance on the American Gas and Electric Company system due to the Norfolk and Western Railway loads will be within safe limits without using other than inherent balance, in preventing damage to generators and utilization apparatus for the combination of conditions that exist on that stretch of railroad.

This is the first attempt in the United States to find out by detailed analysis and actual test the effects of a direct application of an actual single-phase railroad load on an actual commercial-frequency 3-phase line. The results should be very significant.

R. L. Kimball (Gibbs and Hill, Inc., New York, N. Y.): It is a natural reaction during the approach to a new development to anticipate corrective steps which may be taken to offset troubles which might arise. At times these thoughts are found to be inadequate and at others prove to be unnecessary. As pointed out by Mr. Price, the results of taking single-phase loads from a commercial frequency network will be more or less peculiar to each case. However an encouraging start has been made with the careful examination of one instance.

Under the sponsorship of Association of American Railroads Committee 13, a joint quantitative investigation of unbalance was undertaken by the Norfolk and Western Railway, Gibbs & Hill, Inc., and the American Gas and Electric System. Profiles and train data, in the form of dispatchers' train sheets, were supplied by the Norfolk and Western for a representative day of heavy traffic movement over its line from Roanoke to Williamson.

Advantageous points of power supply, both from the supplier's and the user's points of view, were selected by American Gas and Electric and Gibbs and Hill. The latter then converted the train information to kva demands, east and west of each point of supply by 6-minute intervals over the entire 24-hour period. No rearrangements of train departure times or tonnages were made to minimize swings of peak demands. In fact, the only liberties taken were to assume faster road speeds where possible within available locomotive ratings and nonstop movement, at maximum tonnage handled, of a few "set-out" and "pick-up" trains. These changes simplified the calculations and tended to increase rather than minimize power demands.

The resulting data were used by American Gas and Electric in a detailed computer study to determine phase voltage unbalances on its network and phase current unbalances on generating plants now operating or under construction. The entire system south and east of the Philip Sporn plant, including intercompany connections, was included in the basic system data. The results are therefore highly conclusive except for one point. It was not practical to represent the corrective influence of rotating 3-phase loads as it would affect generator current unbalance. This is expected to have a considerably favorable influence.

The demands were arranged on different phases at the various points of supply in order to give some approximation of balance. Random phase selection was made, which later may be found susceptible of improvement. Conditions were calculated on the computer for both straight single-phase and for Scott transformer connections at the points of supply. The following figures refer only to the straight single-phase connections because of insufficient time for a full analysis of results.

Voltage unbalance on the network proved to be much less than anticipated. It can easily be maintained at less than 1%, thereby eliminating worry concerning other consumers' motor equipment. Except for one plant, the generator current unbalance during more or less isolated adverse short-time swings proved not to exceed 2%. This was the result without resorting to the balancing affect of Scott-connected step-down transformers, and also without taking into account the, as yet unevaluated, corrective affect of 3-phase rotating load. Cyclic heating effects are now being evaluated even though the short-time swing values appear probably acceptable.

Still further light on this point will be available from American Gas and Electric generator-rotor tests soon to be undertaken by the General Electric Company.

The indication from these calculations to date is that the effect of voltage unbalance on other loads and of current unbalance on generating plants, in this case, present no problem.

P. A. McGee (Ventnor City, N. J.): The AIEE Committee on Land Transportation owes Mr. Price thanks for his informative and timely paper on power supply for commercial-frequency railroad electrification.

It is a fact that the art of railroad electrification is in a very depressed state and during the past 25 years has shown little, if any, progress with main-line application. It is no wonder then, that the rediscovery of the motor-generator and rectifier locomotives should give new hope to those who, as a matter of principle, or for other reasons, advocate the electrification of a railroad. When it became evident that this type of motive power could possibly work best on commercial frequency rather than on the special low frequency required with a-c series motors, the new hope developed into unrestrained enthusiasm by many advocates. It appeared that possibly the railroads could get something for nothing. The existing transmission facilities could carry perhaps the most complex and costly features of railroad electrification without any cost to the railroad.

This misconception is clearly exposed in the paper's clear presentation of some of the problems involved when commercial frequencies are applied to a single-phase contact system. It appears that in appraising this new system we must consider many factors and follow the good old railroad safety rule at grade crossings: "Stop, Look, and Listen."

The selection of a railroad electrification system, whether through a self-contained diesel or gas-turbine power plant on the locomotive or through a d-c or a-c contact line, must, in a final analysis, depend on the question of economics. Where special conditions such as smoke abatement or strategic or political conditions do not prevail, it is doubtful that any one of the proposed systems of this day can economically supply the various operating conditions for our subways, metropolitan lines, suburban systems, isolated mountain and tunnel operation, and the general network of a large country or continent. On the main-line and suburban tracks of our large railways I find that with electrification supplied with power from a-c or d-c contact systems, there is relatively little difference in the cost of installation and operation of the accepted systems to date. It is further evident that the selection of a particular a-c or d-c contact system in the various countries was largely a matter of local preference or prejudice with the exercise of something less than perfect judgment in many cases.

Although the single-phase motor-generator-type a-c locomotive was developed in a spectacular way for the late Henry Ford on the Detroit, Toledo and Ironton Railroad in the early 1920's and quite an extensive application was made on the Great Northern Railway in the late 1920's, it was not until American railroads had purchased more than 20,000 diesel units and were operating nearly 80% of their systems in 1950 with mass-produced diesel generator locomotives that the similarity of type and many advantages of the diesel generator and the a-c generator locomotives became apparent.

The self-sufficiency and flexibility of the diesel generator to diesel hydraulic motive power, whether applied to locomotives or rail cars, continues, however, to offer an ever-increasing challenge to the motor-generator or rectifier locomotives fed from single-phase contact lines. In the first instance, it should be clearly understood that the electric power required at the traction motor terminals on a locomotive or motor car, whether it is supplied by a diesel engine or gas turbine on the locomotive or through a contact line, is invariably less than half that which must be generated at some central source of power and delivered through substations to an a-c contact line and eventually through the locomotive to the traction motor. When it is further understood that with the continued increase of electric power production in all countries the bulk of this power will be produced by thermal plants of one type or another, the significance of this power conservation with the self-contained unit will at once become apparent. The cost of the transmission lines, substation apparatus, and contact system, however simple, with copper wire at 40 cents a pound, steel at 5 to 6 cents a pound, and labor at $2 to $3 per hour, then presents a capital-outlay and maintenance-cost condition which would at best leave but a small margin of return on the normal financially self-sustaining railroad system.

In regard to the transmission problems discussed by Mr. Price, I would like to refer

to a paper on power supply to main-line railroad contact systems presented in 1933 by E. L. Harder and this discusser.[1] The first part of that paper deals with electrical characteristics and regulation of the contact system. In preparing the paper, we were disappointed to find very definite limits in substation spacing due to poor regulation of the trolley rail circuit with the 25-cycle single-phase system. At 50 or 60 cycles this circuit has very definite substation spacing limits regardless of wire configuration. If the contact sections are fed from only one point, as suggested, to provide against phase shift of the 3-phase supply at the two ends, we would have a highly inductive circuit which, I would think, would be very objectionable to open signal and communication lines or low-voltage power lines running parallel to the tracks.

The necessary dead sections caused by the insulation of the various contact sections introduce a real problem in the frequent breaking of power supply to locomotive during periods of acceleration or on adverse grade. This problem has largely retarded the application of locomotives on third rail systems.

In the second part of the 1933 paper, we show a very interesting condition of power factor over various lengths of track. As an example, a typically loaded U. S. railroad contact line has a load factor of 13 to 22% when the load is supplied at 10-mile intervals. If the load is supplied at 200 100-mile intervals, this load factor may be between 60 and 64%. It will, therefore, be seen that in the application of balancing equipment or, as Mr. Price aptly puts it, a balancing sink, from three to five times the amount of equipment will be required if the power is taken from existing lines without benefit of a separate railway single-phase transmission line.

When one has thoroughly digested the problems presented in the paper and the difficulties which I have outlined, it would appear that there is much to be said for an independent power supply line to a railroad, whether commercial frequencies or special frequencies are employed with an a-c system. There are three large a-c single-phase contact railway systems in the world; the Pennsylvania, with an independent transmission line; the Swedish State Railroad which supplies power to the contact line from motor-generator-phase to single-phase substations taking 3-phase power from the national network; and the Swiss Federal Railroads which have largely their own power plant and independent transmission lines. It is doubtful that any of these extensive systems could in any way be improved economically by changing their power supply to commercial frequencies. One further point I would like to stress is the relative insignificance of a railroad electrification load on our national electric power supply systems as a whole and the abnormal demand of local power if this power were supplied at frequent intervals from existing lines. In this connection, Table IV presents data on the French, Italian, Swedish, and Swiss railroads, indicating the small amount of electric power required even where the railways are almost completely electrified. As these countries progress in the development and use of electric power, it is estimated that within a period of less than 10 years the percentage of national total power required by the railroads will be about half the value shown.

Table IV. Comparative Railroad Electrification Loads

Railroad System	Per Cent of Ton-Miles Hauled Electrically	Railroad Load as Per Cent of National Electric Power Employed
French National	31	2.8
Italian State	64	5.4
Swedish State	87	4.6
Swiss Federal	97	7.0

Reference

1. Power Supply for Main Line Railway Contact Systems, P. A. McGee, E. L. Harder. *AIEE Transactions*, vol. 52, June 1933, pp. 364–78.

H. F. Brown (Gibbs and Hill, Inc., New York, N. Y.): This paper has presented much information which, while not new, is of value for the record because of the diminishing number of engineers in the United States who have had first-hand acquaintance with the problems of single-phase power supply for electrified railroads.

The definition of "commercial frequency electrification" should be a little more specific. The paper deals specifically with the problems of commercial frequency power applied directly to the railroad contact system. Commercial frequency, converted, has been in use by the railways on both a-c and d-c systems for many years.

Two distinct and separate sets of problems are presented in this paper: (1) those which are in the province of the power company to solve; and (2) those which must be solved by the railroads.

In the past, especially prior to 1925, it was not always possible for railroads to purchase commercial electric power for electrification, and nearly all railroads which became electrified had not only to provide their own power supply, but their electrical engineers had to contend with, and solve, many of these problems which were and still are entirely foreign to railway operation.

These were contributing reasons why electrification never was very popular with some railway managements. Any future railway electrification must not present to railway management problems that belong specifically to the power supply companies. This is not to say that the engineers on both sides should not have a clear idea of the mutual problems involved, and the paper points out what some of these problems might be. It does not, however, clearly define the responsibility for providing the solutions.

As an example, consider the matter of large power demands and short-time swings incident to railroad loads. Some of these heavy swings could be reduced by the railroads by changing operating methods if they so desired, which they usually preferred not to do. This made the railroad load undesirable or impossible for some commercial power companies to handle without greatly increasing the power rate, and necessitated, in many cases, the railroad installing its own power supply.

It is safe to say that any future railroad electrification with any type of system will not be made unless the power supply systems in the area of electric operation are of such capacity that they can and will handle railroad load swings of any magnitude presented. This in itself almost immediately rules out the necessity for phase balancing of any residual unbalance of the railroad load offered to the power network. At least it can be said that no phase-balancing equipment has been necessary on systems abroad to supply 50-cycle commercial power directly to the railway contact system. The supply and network for the 50-cycle single-phase electrification of the French National Railways (SNCF) is larger than any single power network in the United States. No phase-balancing equipment has been installed. The 50-cycle network from which the Bas-Congo-Katanga Railway (BCK) in Africa takes its supply is only about 300,000-kw capacity. No phase-balancing equipment has been installed on this supply.

The second largest 25-cycle a-c railroad electrification in the United States takes a large proportion of its single-phase supply directly from the 25-cycle system bus of a large commercial power supply company which is also tied, through many frequency-changer sets, into its 60-cycle network. This 25-cycle system is probably about 600,000 kw in capacity but only three 35,000-kw generators have been equipped with amortisseur windings and there are times when only one of these machines is on the line.

This supply takes all the railroad power swings, which are often 45 megawatts in magnitude, with no appreciable voltage unbalance on the supply station bus. There are no special phase-balancing equipments on the system.

In the railroad-owned power station, equipped with 3-phase generators having amortisseur windings, the load is 95% single phase, and the bus voltages are often unbalanced by as much as 5%. All of the station auxiliaries are normally electrically driven by 25-cycle squirrel-cage-type induction motors of standard design supplied from the station bus. They run "hot," due to the unbalance, but not to the extent that deterioration of the windings has occurred. During the early years of operation when street railways were in vogue, 25-cycle rotary converters were successfully operated from this unbalanced system.

It is recognized that the 60-cycle problem will be slightly different from the 25-cycle or 50-cycle problem in this detail, but it is also true that the commercial power engineers were somewhat sceptical in advance concerning the effects of single-phase railway loads, both in the United States and abroad, before the loads were actually applied. After careful tests in all the aforementioned cases, the power engineers agreed that "there is no problem."

I foresee no phase-balancing equipment required in any event on the part of the railroads. The railroads, and their engineers, must, nevertheless recognize their share of the problems. However, the cost should be comparatively little for the solutions and practically nothing for operation.

1. The lines must be sectionalized at strategic locations to permit such inherent balancing of the load on the different phases

of the network as may be practicable. This introduces no new problem for railway electrification as nearly all electrified a-c lines are now sectionalized at track interlocking locations, and phase breaks are not unknown on the existing 25-cycle a-c systems.

2. Just as the existing d-c electrified railroads have had to co-operate with power supply companies in dealing with harmonics from their stationary rectifier substations, so will any future single-phase a-c railroad have to do all that can be done in this direction by the installation of adequate filters or harmonic shunts on their moving rectifier equipment. The problem will, to some degree, be simpler because of the smaller capacity and greater number of the rectifier units than in fixed substations of large capacity, and will also provide some incidental power-factor improvements.

3. The major problem of the railroads will be that of reducing to minimum values the inductive disturbances inherent in all single-phase grounded a-c systems to enable them to exist amicably with other neighboring utilities providing communications. Here the wealth of experience already gained on 25-cycle installations will prove valuable, and must not be ignored in designing new methods of power supply and distribution. These solutions are well known to the engineers of the communication utilities, and also some of the railroad electrical engineers.

This problem is given but brief mention in the paper as being outside the scope, although its consideration would have certainly vetoed some of the distribution methods proposed and given considerable emphasis in the paper.

The methods of distribution outlined in Figs. 1(A), 1(C), and 1(D) would all be questionable from the standpoint of inductive disturbances. The method shown in Fig. 3 would be acceptable from the inductive standpoint, but inacceptable from the railroad's operating and maintenance standpoint.

The methods of distribution satisfactory from the railroad point of view shown in the paper are Figs. 1(B) and 2.

As already stated, I do not visualize the necessity, as part of the railroad substation equipment, of such schemes as shown in Figs. 4 and 5. Such schemes imply that railroad loads are supplied from inadequate power sources.

Of course special applications of railway electrifications, especially in foreign countries where fuel is scarce, or in undeveloped areas, may indicate special phase-balancing apparatus as pointed out in an AIEE paper by this discusser and R. L. Kimball in 1954.[1] I used the method shown in Fig. 6 for a proposed commercial-frequency electrification in South America in 1953, where the railroad load was one half that of the total power supply.

I believe that the 3-wire system, which is given but scant mention, will be found especially helpful in solving both distribution and inductive co-ordination problems, in any future application of 60-cycle single-phase alternating current to the railways, just as it has been in existing 25-cycle installations in the United States. This feature alone is worth considerably more study than has been accorded it in most of the recent studies of 60-cycle supply for railroad electrification.

As a supplement to the appendix showing resolution of single-phase loads into 3-phase sequence components, and as an addition to the references in the paper, attention is called to a paper by J. Daycard of the Chilean State Electric Power System (ENDESA).[2]

This study was made to determine the effect on the existing power network of electrifying the Chilean State Railway south of Santiago with single-phase commercial frequency directly on the railway contact line following a railroad electrification study based on the use of commercial single-phase a-c power. This system is not too large, about 300,000 kw, and is mostly water power. The study showed that "the proposed loads would be accepted either for direct connection or with the addition of some kind of phase balancers."

Mr. Price is to be complimented on the scope and treatment of the problems in this paper. The comments offered herewith are not to be construed as criticism of the paper. They are offered as sidelights on the problem from the railroad standpoint by one who, after spending the better part of a lifetime in dealing with many of these problems on a railroad successfully electrified using the 25-cycle a-c single-phase system, is convinced that such a system using any commercial frequency offers the most economic method for any future large-scale railway electrification.

References

1. A Reappraisal of the Economics of Railway Electrification: How, When, and Where Can It Compete with the Diesel-Electric Locomotive? H. F. Brown, R. L. Kimball. *AIEE Transactions*, vol. 73, pt. II, Mar. 1954, pp. 33–51.

2. Network Analyzer Studies of Unbalanced Conditions Arising from Single-Phase Loads in a Three-Phase Power System, J. Daycard, *Paper no. 301*, CIGRE, Paris, France, 1956.

J. C. Price: The discussers of this paper have made available to those interested in railroad electrification much valuable information based upon their many years of first-hand experience in this field. Some of the discussion is outside of the subject matter encompassed by the paper, and in these areas I will comment upon discussion to the extent that my experience permits me to do so.

I appreciate the descriptions by Mr. Curtis and Mr. Kimball of the pioneering work undertaken in the field of commercial-frequency electrification under the sponsorship of Committee 13 of the Association of American Railroads. The activity stimulated by this committee promises to develop significant additions to the storehouse of practical knowledge concerning commercial-frequency electrification.

I share Mr. Curtis' conviction that electrification of heavily traveled main track will merit consideration as increasing demand upon the petroleum supplies available to the United States exerts pressure on the price of locomotive fuel, and as maintenance costs of internal-combustion engines become more thoroughly evaluated. Further reason for optimism concerning the future for railroad electrification may be found in the trend toward improved utilization of track through traffic control, traffic growth, and consolidation of operations; and in the need for increased locomotive speed and horsepower to improve the competitive position of rail carriers. Electrification not only benefits in economic analysis from increased track utilization and increased speed of operation, but it is, in fact, better able in a technical sense to make possible higher levels of performance in these areas than other presently available systems of motive power.

Mr. McGee's comment concerning the greater prime generation requirement for electric operation, as compared with diesel operation, calls attention to the potential economy that electrification offers from a fuel viewpoint. This apparent contradiction is explained by consideration of the fuel economy of central-station generation as contrasted with the relatively high fuel cost for generating power in small isolated power plants.

At the level of diesel oil prices prevailing in the United States today, about 11 cents per gallon, the fuel cost for each kilowatt-hour generated on board a diesel locomotive is of the order of 10 mills.[1] Electrical utility generating stations in 1954 showed an average fuel cost of 3 mills per kilowatt-hour generated, on a nation-wide basis.[2] This figure was heavily weighted by generation costs in coal-burning plants which accounted for approximately two thirds of the total generation reported during that year.

Thus, fuel economy, measured in dollars, promises to favor electrification even if we accept the proposition that electrification requires the prime generation of twice as many kilowatt-hours as does diesel operation. On this basis, fuel costing 6 mills when burned in a typical generating station will do the same work (and, as will be seen, at a faster rate) as will be done by fuel costing 10 mills when burned in an isolated power plant aboard a diesel locomotive.

As Mr. McGee emphasizes, the prospects for electrification depend, in the final analysis, upon over-all economic considerations, and not upon fuel costs alone. In many areas outside of the United States, the fuel situation is vastly different than it is here, and this, combined with special conditions of a local nature, may indicate electrification as a preferred system of operation in those areas. Given the present fortunate position of the United States with respect to availability of petroleum, the over-all economics of the situation can be expected to prevail, and widespread electrification of main-line railroad is hardly a near-term prospect here. This does not, however, exclude the desirability of continued study of electrification by certain U. S. railroads which lead the industry in traffic density and in the need for high performance of motive power.

The relative prime generation requirements for electric operation and for diesel operation deserve further explanation. Where comparison is based upon equivalent performance, it is reasonable to expect up to 50% more kilowatt-hours of prime generation to support an electric operation than to support a diesel one. Take, for example, the case of motor-generator electric locomotives which are comparable in performance with diesel-powered locomotives if both have the same horsepower rating referred to the generator input. The following efficiencies are reasonable for this case, and will serve for illustration:

Motor-generator locomotive (no regeneration)	0.74
25-kv distribution	0.95
Traction step-down substation	0.96
Electrical utility system transformation and transmission	0.92
Over-all efficiency	0.62

In the case of the diesel locomotive, traction motor efficiency may be taken as approximately 0.92.

The ratio of kilowatt-hours generated in this particular electric system and in the diesel system thus becomes $\frac{0.92}{0.62} = 1.48$. This indicates 50% more kilowatt-hour generation required for the over-all motor-generator system than for the diesel system.

If the electric system is one of higher performance than the diesel system, then the ratio of kilowatt-hours between the two systems may exceed this value, because kilowatt-hours per unit of traffic (gross ton-miles or passenger-car miles) increases as the speed of an operation increases. To the extent that prime energy requirements for electrification may extend beyond a 50% increase over diesel requirements, this is a measure of superior performance of electric motive power.

To introduce the next phase of discussion, it is desirable to state the definition of commercial-frequency electrification as used for purposes of this paper. I regarded commercial-frequency electrification as electrification which employs on the contact system, power from the same prime source which is available to serve industrial, commercial, and residential loads in an area; and which employs this power without the use of intermediate energy conversion or conversion of a nature that changes the frequency of power applied to the contact system. Under this definition, none of the major 25-cycle electrifications which have existed in the United States would qualify as a commercial-frequency electrification. Mr. Brown correctly points out, however, that large single-phase traction loads have been taken directly from an extensive 25-cycle network. I regard this situation and commercial-frequency electrification as being analogous, but not identical, because the 25-cycle network has supplied principally special-purpose loads.

The advice to "stop, look, and listen" before regarding "commercial-frequency" as the answer to all electrification problems is sound. Committee 13 is to be commended in this connection for initiating a study which will reveal the practicability of commercial-frequency electrification load from a power-supply viewpoint in one specific case. In a recent paper[5] we see evidence that the matter of single-phase loading of a power system has received detailed attention in Chile. Papers from French, British, and other European sources provide evidence that the effects of single-phase load have been studied in connection with 50-cycle electrifications in operation or proposed for several European countries.

It appears reasonable to expect, however, that no insurmountable technical problems will be encountered when considering commercial-frequency electrification in the United States. Engineering decisions will have to be made in many areas and the discussers have called attention to a number of these areas.

As Mr. Curtis points out, the relatively simple power transmission and distribution arrangements shown in Fig. 1 will not always be realized in practice. In areas with widespread industrial development, the general case of commercial-frequency electrification may well be one which involves localized unbalanced loading at several points on one or more power systems. The effect of unbalanced traction load upon near-by loads connected to the electric supply system, and the effect upon near-by sources of generation must be evaluated.

Where a commercial-frequency electrification is constructed through an area with relatively undeveloped power facilities, it may be necessary to parallel the railroad with a power transmission line. This has at times been necessary in order to electrify at 3,000 volts d-c, both in the United States and abroad, and the need for transmission facilities would not be altered by the fact even though a different system of electrification were selected.

One major expectation, of course, is that commercial-frequency electrification will make unnecessary an independent supply line for an a-c railroad electrification. Yet, it is conceivable that in an individual case there may be economic justification for the independent supply line, as Mr. McGee indicates. For example, should phase balancing be a prime consideration in a particular case, it may prove more economical to accomplish balancing at a single, central point than to balance at a large number of power delivery points. This can be done by combining the loads of a number of traction substations through a transmission line, and taking delivery of power at a single point. This would be a contributing factor to the case for an independent transmission line. Other contributing factors can be visualized: the possible economy of purchasing power at a single point, the greater freedom which may be available in providing economical protection for substation equipment operating at primary voltage, and possible elimination of end-fed arrangements necessitated by phase displacements between voltages at different points on power systems.

Trolley-rail regulation, which was mentioned by Mr. McGee, is a consideration in planning power supply for any system of railroad electrification. In connection with commercial-frequency electrification in France, detailed analysis showed improved voltage conditions for distribution of a given number of kilowatts when considering 50-cycle 25-kv electrification instead of $16^{2}/_{3}$-cycle 15-kv electrification, even under relatively low power-factor conditions.[4] The comparison between 60-cycle 25-kv electrification and 25-cycle 11-kv electrification is even more favorable, because of the relatively greater voltage increase and the relatively smaller frequency increase in this case.

At unity power factor, it can be shown that the decrease in voltage at a locomotive operating at 25 kv, 60 cycles, is relatively small, even for large values of load. (The IZ drop in the system may be large, but it leads the locomotive voltage by perhaps 70 degrees so that the difference between scalar values of substation voltage and locomotive voltage does not greatly exceed the IR drop). This suggests that shunt capacitors may find useful application in correcting power factor on commercial-frequency electrifications where substation spacing is a problem. Series capacitors may also find application in such situations as a means of partially compensating for contact system reactance. There is precedence in a-c electrification practice for the use of both shunt capacitors[3] and series capacitors.[3]

I agree with Mr. Brown and Mr. McGee that consideration of communications interference will be an important factor in planning commercial-frequency electrification. The operation of 25-cycle rectifier locomotives in the United States has provided interesting insight into the nature of problems which may be encountered with this type of motive power operating from a 60-cycle supply.[7] Committee 13, in connection with its study of commercial-frequency electrification, has devoted attention to this problem.[8] Various papers presented at Annecy in 1951 and at Lille in 1955 in connection with French developments in commercial-frequency electrification explore this matter in detail. In this area, as in the areas specifically treated in the paper, the particular conditions applying to a railroad which is being considered for electrification will provide the final answer concerning system design features and remedial measures required because of communications interference considerations.

References

1. See reference 1 of Mr. Brown's discussion.

2. Electric Industry Statistics. *Electrical World*, New York, N. Y., Jan. 23, 1956.

3. See reference 2 of Mr. Brown's discussion.

4. Voltage Drops in the Contact Wires at 50 Hz Single-Phase Traction, M. Bernard. Presented at the Convention on Electric Traction by Single-Phase Current at Industrial Frequency, Annecy, France, 1951.

5. Shunt Capacitor Installation for Single-Phase Railway Service, H. F. Brown, R. L. Witzke. *AIEE Transactions*, vol. 67, pt. I, 1948, pp. 258–260.

6. Capacitor Stations Installed in Series on the Aix-Les-Bains to Annecy Experimental Line, M. Barboutan. Presented at the Convention on Electric Traction by Single-Phase Current at Industrial Frequency, Annecy, France, 1951.

7. Rectifier Motive Power—Inductive Coordination Considerations, E. B. King, K. H. Gordon, L. J. Hibbard. *AIEE Transactions*, vol. 73, pt. II, July 1954, pp. 107–18.

8. Report of Committee No. 13—Railway Electrification Assignment #5, Task Group 3 on Inductive Interference. Advance Reports of the Fourth Annual Meeting, Association of American Railroads, Electrical Section of the Engineering and Mechanical Divisions, 1956